Cognadev Technical Report Series

6

17th June, 2016

Hierarchical Multiple Linear Regression and the correct interpretation of the magnitude of a Deviation R-square (ΔR^2).

I read article after article where psychologists interpret what look to me to be trivial ΔR^2 values as though they were meaningful. Either my judgement is deeply flawed, or the judgement of the authors who report trivial ΔR^2 values as meaningful is flawed.

So, in this Technical Report, I seek the answer to two questions:

❶ Does a small ΔR^2 value have any pragmatic value at all?

❷ What magnitude of ΔR^2 is worth reporting beyond: *nothing to see here* ?

Contents

Tables

Figures

1. Hierarchical Multiple Linear Regression

In hierarchical linear regression, models are fitted to a dataset predicting a single outcome variable *(usually)*; where each model is constructed by adding variables to an initial equation, and computing a deviation R-square (ΔR^2) which is the difference between an initial model *(or previous model in the sequence)* R^2 and the new model R^2. This might be done 3 or 4 times, as blocks of variables are added incrementally to an initial block, and their impact assessed on predictive accuracy using the ΔR^2 magnitudes.

For example, a researcher might be interested in the incremental predictive accuracy gained from initially predicting job-performance using 2 ability variables, then the extra accuracy created by including 3 personality, and then 2 motivation variables to predict the same job-performance.

Model 1 - ability

$$jp = constant + b_1 * a_1 + b_2 * a_2$$

$$\text{model } R\text{-squared} = R^2_{m1}$$

Model 2 - ability + personality

$$jp = constant + b_1 * a_1 + b_2 * a_2 + b_3 * p_1 + b_4 * p_2 + b_5 * p_3$$

$$\text{model } R\text{-squared} = R^2_{m2} \quad \text{with } \Delta R^2_{m2-m1} = R^2_{m2} - R^2_{m1}$$

Model 2 - ability + personality + motivation

$$jp = constant + b_1 * a_1 + b_2 * a_2 + b_3 * p_1 + b_4 * p_2 + b_5 * p_3 + b_6 * m_1 + b_7 * m_2$$

$$\text{model } R\text{-squared} = R^2_{m3} \quad \text{with } \Delta R^2_{m3-m2} = R^2_{m3} - R^2_{m2}$$

Conventionally, each model's incremental fit (R^2) over the previous model is tested for statistical significance. This is implemented using an ANOVA[1] approach

$$F^H_{n-K} = \frac{RSS_{smaller} - \left(\frac{RSS_{\text{larger}}}{H} \right)}{\left(\frac{RSS_{\text{larger}}}{(n-K)} \right)}$$

where

F^H_{n-K} = F distribution statistic with H and $(n - K)$ d.f.

$RSS_{smaller}$ = the residual sum of squares for the fewer parameters regression model

RSS_{larger} = the residual sum of squares for the greater no. of parameters regression model

H = the number of parameters for the smaller (fewer parameters) model

K = the number or parameters for the larger (greater no. of parameters) model

n = the total number of cases

[1] Hamilton, L.C. (1992) *Regression with Graphics: A Second Course in Applied Statistics*. Belmont, California: Brooks-Cole (see Eq. 3.28, page 80)

1.1 Adjusted R^2

In any multiple regression situation, the model R^2 is adjusted/corrected for the upward bias in the estimate due to capitalisation on chance as a result of the number of predictors in an equation. The correction formula and a worked example is:

$Rsq := 0.28$	model R-square
$R := \sqrt{Rsq} = 0.52915$	observed multiple R
$m := 12$	number of predictors
$N := 514$	number of cases
$R_{adj} := R^2 - (1 - R^2)\left(\frac{m}{(N - m - 1)}\right) = 0.263$	unbiased/adjusted multiple R

It's an important and sometimes substantive correction *(depending upon the number of predictors and sample size)*.

Question. Should the ΔR^2 be computed using model R^2 or the *adjusted* ΔR^2?

Answer. Given the logic of the correction, it only makes sense to compute the ΔR^2 using the adjusted ΔR^2, as this is the best unbiased estimate of predictive accuracy.

1.2 Two Questions

Many researchers seem quite happy to use a statistically significant ΔR^2 as low as 0.01 as 'evidence' for an incremental effect, which in the "Discussion" or "Conclusions" to an article invariably ends up with a statement of the form: *"a significant incremental effect of attribute X was observed over and above Y and Z, indicating that attribute X is worth considering alongside Y and Z."*

Personally, I think this is deeply flawed. But does the flaw reside in my judgement or in that of the authors who choose to report such small increments as meaningful?

I want an answer to two questions:

❶ Does such a small ΔR^2 value have any pragmatic value at all?

❷ What magnitude of ΔR^2 is worth reporting as more than *"nothing to see here"*?

To answer question ❶, I'm going to use the correlation matrix from a published study and from it, generate the raw data which would create such a matrix.

The published study was sent to me by a student who, like the small boy in the Emperor's New Clothes fable, simply couldn't see how the claim of *'important effect'* made by the authors could ever be substantiated by the tiny ΔR^2 they reported.

There is nothing personal here; the article is simply a good exemplar of all such articles (*and student theses*) which proudly present what looks to be *'nothing to see here'* as 'substance'.

To answer question ❷, I'm going to simulate integer-score data showing a .29 and .31 correlational relationship, in order to get a feel for what magnitude of R^2 might be seen as 'useful' in terms of predictive accuracy.

2. The example published study

Akhtar, R., Boustani, L., Tsivrikos, D., & Chamorro-Premuzic, T. (2015). The engageable personality: Personality and trait EI as predictors of work engagement. *Personality and Individual Differences*, 73, 44-49.

Abstract

Work engagement is seen as a critical antecedent of various organizational outcomes such as citizenship behavior and employee productivity. Though defined as a state, recent research has hinted at potential individual differences in engagement, meaning that employees differ in their tendencies to engage at work. This study investigated the effects of the Big Five personality traits, work-specific personality, and trait emotional intelligence, on work engagement among a sample of 1050 working adults. Hierarchical multiple regression analyses identified trait EI, openness to experience, interpersonal sensitivity, ambition, extraversion, adjustment, and conscientiousness as predictors of engagement. Trait EI predicted work engagement over and above personality. Practical and theoretical implications are discussed.

Table 1: The four regression model variables and parameters within in the published article (Table 2 in that article)

	Model 1		Model 2		Model 3		Model 4	
Variables	β	t	β	t	β	t	β	t
Age	.24	7.99***	.18	6.24***	.16	5.70***	.16	5.62***
Gender	.01	.27	−.01	−.49	−.01	−.18	−.02	−.51
Neuroticism			−.12	−3.79***	−.02	−.51	.03	.74
Conscientiousness			.14	4.99***	.10	3.45**	.08	2.46*
Openness			.20	6.90***	.16	5.31***	.13	4.15***
Agreeableness			−.01	−.16	−.05	−1.57	−.07	−1.09*
Extraversion			.18	6.19***	.13	3.55***	.11	2.86**
Sociability					.04	1.08	.02	.54
Interpersonal sensitivity					.14	4.34***	.12	3.85***
Prudence					.04	1.29	.04	1.16
Inquisitive					−.01	−.27	−.01	−.21
Adjustment					.12	3.36**	.09	2.31*
Ambition					.14	4.84***	.12	4.20***
Learning approach					.06	1.92	.04	1.22
Trait EI							.18	4.16***

	$R2$	Adj R^2	R^2 Change
Model 1	.058	.056	.058
Model 2	.203	.197	.145
Model 3	.255	.245	.052
Model 4	.267	.256	.012

Note also that the *unadjusted* R^2 values are used to compute the "Change" values - so inflating them.

* Correlation significant at the .05 level (2-tailed).
** Correlation significant at the .01 level (2-tailed).
*** Correlation significant at the .01 level (2-tailed). $N = 1050$.

On the basis of these results, the authors state:
"Our results provide an insightful prospective towards a hierarchical integration of dispositional determinants for work engagement, especially highlighting the independent contribution of trait EI in the prediction of engagement. Broad measures of personality, along with work-specific measures and trait EI appear to be important contributors to work engagement." (p, 48, column 1, 3rd para)

Table 2: The correlation matrix between study variables (screenshot from the article)

Table 1
Descriptive statistics & bivariate correlations.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
1. Engagement	–														
2. E	.24**	–													
3. A	.12**	.03	–												
4. C	.20**	−.01	.15**	–											
5. N	−.20**	−.05	−.36**	−.23**	–										
6. O	.31**	.28**	.13**	.09**	−.19**	–									
7. Adjustment	.27**	.06	.21**	.17**	−.65**	.19**	–								
8. Ambition	.25**	.22**	−.08**	.12**	−.12	.19**	.28**	–							
9. Sociability	.21**	.67**	.06	−.07*	−.15	.28**	.31**	.06	–						
10. IS	.18**	.01	.53**	.13**	−.23**	.12**	.31**	.15**	−.05	–					
11. Prudence	.07*	-.07*	.10**	.38**	−.10**	−.13**	.11**	.08**	.05	−.14**	–				
12. Inquisitive	.02	.13**	−.11**	−.11**	.11**	.22**	.02	−.04	.13**	.17**	−.08*	–			
13. LA	.19**	.05	.07	.08*	−.08**	.29**	.26**	.12**	.15**	.11**	.15**	−.01	–		
14. Trait EI	.41**	.33**	.36**	.33**	−.57**	.43**	.52**	.17**	.22**	.05	.13**	−.28**	.18**	–	
15. Age	.24**	.00	.18**	.09**	−.20**	.11**	.21**	.02	.07*	.09**	.00	−.15**	.16**	.21**	–
Mean	4.46	4.63	4.96	5.56	3.02	5.68	3.15	3.23	3.28	3.90	3.37	2.92	3.95	3.85	45.29
SD	.91	1.37	1.11	1.12	1.27	1.00	.81	.69	.80	.64	.74	.66	.62	.44	12.47
α	.90	.63	.25	.47	.57	.45	.67	.37	.56	.37	.49	.20	.28	.88	–

Note: E = Extraversion, C = Conscientiousness, A = Agreeableness, O = Openness, N = Neuroticism, IS = Interpersonal Sensitivity, LA = Learning Approach.
* Correlation significant at the .05 level (2-tailed).
** Correlation significant at the .01 level (2-tailed).

But note that "Gender" which appears as a prediction variable in the Hierarchical models (see article Table 2 above) does not appear in this matrix. We will also ignore the alpha reliabilities as low as 0.20 … So, next step was to enter the correlation matrix into Statistica, in readiness for the analysis.

Table 3: The Statistica transcribed correlation matrix from Akhtar et al (2016)

	Akhtar et al correlation matrix														
	1 Engagement	2 Extraversion	3 Agreeable	4 Conscientious	5 Neuroticism	6 Open to Experience	7 Adjustment	8 Ambition	9 Sociability	10 Interpersonal Sensitivity	11 Prudence	12 Inquisitive	13 Learning Approach	14 Trait EI	15 Age
Engagement	1	0.24	0.12	0.2	-0.2	0.31	0.27	0.25	0.21	0.18	0.07	0.02	0.19	0.41	0.24
Extraversion	0.24	1	0.03	-0.01	-0.05	0.28	0.06	0.22	0.67	0.01	-0.07	0.13	0.05	0.33	0
Agreeable	0.12	0.03	1	0.15	-0.36	0.13	0.21	-0.08	0.06	0.53	0.1	-0.11	0.07	0.36	0.18
Conscientious	0.2	-0.01	0.15	1	-0.23	0.09	0.17	0.12	-0.07	0.13	0.38	-0.11	0.08	0.33	0.09
Neuroticism	-0.2	-0.05	-0.36	-0.23	1	-0.19	-0.65	-0.12	-0.15	-0.23	-0.1	0.11	-0.08	-0.57	-0.2
Open-to-Experience	0.31	0.28	0.13	0.09	-0.19	1	0.19	0.19	0.28	0.12	-0.13	0.22	0.29	0.43	0.11
Adjustment	0.27	0.06	0.21	0.17	-0.65	0.19	1	0.28	0.31	0.31	0.11	0.02	0.26	0.52	0.21
Ambition	0.25	0.22	-0.08	0.12	-0.12	0.19	0.28	1	0.06	0.15	0.08	-0.04	0.12	0.17	0.02
Sociability	0.21	0.67	0.06	-0.07	-0.15	0.28	0.31	0.06	1	-0.05	0.05	0.13	0.15	0.22	0.07
Interpersonal Sensitivity	0.18	0.01	0.53	0.13	-0.23	0.12	0.31	0.15	-0.05	1	-0.14	0.17	0.11	0.05	0.09
Prudence	0.07	-0.07	0.1	0.38	-0.1	-0.13	0.11	0.08	0.05	-0.14	1	-0.08	0.15	0.13	0
Inquisitive	0.02	0.13	-0.11	-0.11	0.11	0.22	0.02	-0.04	0.13	0.17	-0.08	1	-0.01	-0.28	-0.15
Learning Approach	0.19	0.05	0.07	0.08	-0.08	0.29	0.26	0.12	0.15	0.11	0.15	-0.01	1	0.18	0.16
Trait EI	0.41	0.33	0.36	0.33	-0.57	0.43	0.52	0.17	0.22	0.05	0.13	-0.28	0.18	1	0.21
Age	0.24	0	0.18	0.09	-0.2	0.11	0.21	0.02	0.07	0.09	0	-0.15	0.16	0.21	1
Means	4.46	4.63	4.96	5.56	3.02	5.68	3.15	3.23	3.28	3.9	3.37	2.92	3.95	3.85	45.29
Std.Dev.	0.91	1.37	1.11	1.12	1.27	1	0.81	0.69	0.8	0.64	0.74	0.66	0.62	0.44	12.47
No.Cases	1050														
Matrix	1														

Next, I recomputed all the Model regression statistics using this transcribed correlation matrix, to gauge the degree of error incurred because of the rounding to two decimal places of all correlation coefficients (as well as the impact of the missing Gender variable which was not reported in the article correlation matrix).

Table 4: Transcribed correlation matrix check: Regression model results compared to those of Akhtar et al

	Akhtar et al values				**Transcribed correlation matrix values**			
	R^2	Adjusted R^2	Unadjusted ΔR^2	Adjusted ΔR^2	R^2	Adjusted R^2	Unadjusted ΔR^2	Adjusted ΔR^2
Model 1	.058	.056	.058	.056	.0576	.0567	.058	.058
Model 2	.203	.197	.145	.141	.1987	.1941	.141	.137
Model 3	.255	.245	.052	.048	.2395	.2300	.041	.036
Model 4	.267	.256	.012	.011	.3024	.2930	.063	.063

Not quite the same, but 'good enough' given the missing gender variable and rounding to two decimal places for the input matrix. However, what we really need is the raw data from which these correlations were generated. Rather than having to pester the authors for their data, it is possible to generate raw data which conforms to the observed

sample means and standard deviations, and which will reproduce the observed correlation matrix. 1050 cases of such data were generated using the Statistica Data Simulation module, where every variable is assumed to be normally distributed, with mean and SD as per published values, and minimum and maximum-possible value constraints applied to each variable. The data generation method chosen was Latin Hypercube Sampling with Iman Conover preservation of the rank-order structure of correlations in the observed correlation matrix. I'm retaining the real-valued data estimates as the authors express every integer sum-scale score as a fraction of the number of items in a scale rather than preserve the integer data metrics.

We need the raw data because I want to compare our observed outcome variable (the Engagement scores) with their predicted equivalents provided by each regression model fit to them. In this way, we get to see the actual impact of ΔR^2 values in the metric of the observed variable whose 'variation' supposedly being accounted for,

As a check on the success of the data generation (in terms of reproducing the correlations, and means and SDs), the difference between the published and computed matrix (using the generated data) is presented in Table 5.

Table 5: The deviations between the correlations in Akhtar et al (2016) minus the correlations computed from the generated raw data

	Signed differences between Akhtar et al correlations and the correlations from the generated data														
	1 Engagement	2 Extraversion	3 Agreeable	4 Conscientious	5 Neuroticism	6 Open to Experience	7 Adjustment	8 Ambition	9 Sociability	10 Interpersonal Sensitivity	11 Prudence	12 Inquisitive	13 Learning Approach	14 Trait EI	15 Age
Engagement		.001	.001	.007	-.002	.003	.002	.005	.004	.001	.001	.002	.004	.002	.002
Extraversion	.001		-.004	-.003	.003	.007	-.0	.006	.006	-.001	.003	.002	-.001	.006	.002
Agreeable	.001	-.004		-.002	-.007	-.002	.004	.002	.002	-.001	.004	-.008	.007	.005	.001
Conscientious	.007	-.003	-.002		.001	.004	.006	.010	-.007	-.003	.011	.0	.006	.001	.004
Neuroticism	-.002	.003	-.007	.001		.006	-.005	.002	.006	-.002	-.0	.002	-.008	-.004	-.009
Open to Experience	.003	.007	-.002	.004	.006		-.003	.011	.002	.001	-.003	.003	.004	.002	-.0
Adjustment	.002	-.0	.004	.006	-.005	-.003		-.003	-.002	.003	-.001	.002	.005	-.0	-.001
Ambition	.005	.006	.002	.010	.002	.011	-.003		-.001	.004	-.003	.002	.004	.003	.001
Sociability	.004	.006	.002	-.007	.006	.002	-.002	-.001		.002	-.0	.006	.002	-.002	-.002
Interpersonal Sensitivity	.001	-.001	-.001	-.003	-.002	.001	.003	.004	.002		-.007	-.004	.005	.003	-.001
Prudence	.001	.003	.004	.011	-.0	-.003	-.001	-.003	-.0	-.007		-.0	-.004	-.003	.0
Inquisitive	.002	.002	-.008	.0	.002	.003	.002	.002	.006	-.004	-.0		-.006	.002	-.006
Learning Approach	.004	-.001	.007	.006	-.008	.004	.005	.004	.002	.005	-.004	-.006		.0	.003
Trait EI	.002	.006	.005	.001	-.004	.002	-.0	.003	-.002	.003	-.003	.002	.0		-.001
Age	.002	.002	.001	.004	-.009	-.0	-.001	.001	-.002	-.001	.0	-.006	.003	-.001	
Means	.077	.108	.068	.20	-.133	.169	.012	.0	.020	.051	.016	.0	.053	.0	-.367
Std.Dev.	.104	.150	.106	.199	.163	.172	.044	.026	.052	.071	.044	.025	.071	.017	.848
No.Cases	1050														
Matrix	1														

The result indicates the differences are trivial. So, now we compute the regression models and investigate the impact of the ΔR^2 values on the prediction of our outcome variable scores for Engagement.

3. Predicting Engagement Scores

3.1 The regression models

Using the simulated raw dataset, n=1050 cases; the published models (from Table 2 in the article, Table 1 above) were fitted to the simulated dataset.

Table 6: Regression parameters for Model 1

N=1050	**Model 1** Regression Summary for Dependent Variable: Engagement R= .23812887 R²= .05670536 Adjusted R²= .05580527 F(1,1048)=63.000 p<.00000 Std.Error of estimate: .78312					
	beta	Std.Err. of beta	b	Std.Err. of b	t(1048)	p-value
Intercept			3.63	0.10	37.0	0.00
Age	0.24	0.03	0.02	0.00	7.9	0.00

Table 7: Regression parameters for Model 2

N=1050	**Model 2** Regression Summary for Dependent Variable: Engagement R= .44161047 R²= .19501980 Adjusted R²= .19038905 F(6,1043)=42.114 p<0.0000 Std.Error of estimate: .72516					
	beta	Std.Err. of beta	b	Std.Err. of b	t(1043)	p-value
Intercept			1.63	0.27	6.0	0.00
Age	0.19	0.03	0.01	0.00	6.6	0.00
Neuroticism	-0.08	0.03	-0.06	0.02	-2.5	0.01
Conscientious	0.14	0.03	0.12	0.03	4.9	0.00
Open to Experience	0.21	0.03	0.20	0.03	7.1	0.00
Agreeable	0.00	0.03	0.00	0.02	0.1	0.92
Extraversion	0.18	0.03	0.12	0.02	6.2	0.00

Table 8: Regression parameters for Model 3

N=1050	**Model 3** Regression Summary for Dependent Variable: Engagement R= .48589326 R²= .23609226 Adjusted R²= .22650655 F(13,1036)=24.630 p<0.0000 Std.Error of estimate: .70880					
	beta	Std.Err. of beta	b	Std.Err. of b	t(1036)	p-value
Intercept			0.29	0.34	0.86	0.39
Age	0.17	0.03	0.01	0.00	6.09	0.00
Neuroticism	0.01	0.04	0.00	0.03	0.16	0.87
Conscientious	0.10	0.03	0.09	0.03	3.21	0.00
Open to Experience	0.19	0.03	0.19	0.03	5.92	0.00
Agreeable	-0.04	0.04	-0.03	0.03	-0.99	0.32
Extraversion	0.16	0.04	0.11	0.03	4.03	0.00
Sociability	0.00	0.04	0.00	0.05	0.10	0.92
Interpersonal Sensitivity	0.11	0.04	0.16	0.05	2.89	0.00
Prudence	0.06	0.03	0.07	0.04	1.85	0.06
Inquisitive	-0.02	0.03	-0.03	0.04	-0.79	0.43
Adjustment	0.10	0.04	0.11	0.05	2.33	0.02
Ambition	0.11	0.03	0.13	0.04	3.32	0.00
Learning Approach	0.03	0.03	0.05	0.04	1.02	0.31

Table 9: Regression parameters for Model 4

N=1050	**Model 4** Regression Summary for Dependent Variable: Engagement R= .54334000 R^2= .29521836 Adjusted R^2= .28568508 F(14,1035)=30.967 p<0.0000 Std.Error of estimate: .68115					
	beta	Std.Err. of beta	b	Std.Err. of b	t(1035)	p-value
Intercept			-2.2	0.42	-5.2	0.00
Age	0.17	0.03	0.0	0.00	6.3	0.00
Neuroticism	0.07	0.04	0.0	0.03	1.8	0.07
Conscientious	0.04	0.03	0.0	0.03	1.2	0.25
Open to Experience	0.02	0.04	0.0	0.04	0.6	0.52
Agreeable	-0.16	0.04	-0.1	0.03	-4.2	0.00
Extraversion	-0.07	0.05	-0.0	0.03	-1.4	0.15
Sociability	0.17	0.04	0.2	0.05	3.8	0.00
Interpersonal Sensitivity	0.24	0.04	0.3	0.06	6.2	0.00
Prudence	0.04	0.03	0.0	0.04	1.3	0.19
Inquisitive	0.11	0.03	0.1	0.04	3.4	0.00
Adjustment	-0.14	0.05	-0.1	0.05	-2.8	0.01
Ambition	0.15	0.03	0.2	0.04	4.9	0.00
Learning Approach	0.04	0.03	0.1	0.04	1.4	0.15
Trait EI	0.48	0.05	0.9	0.10	9.3	0.00

Table 10: The Akhtar et al and generated raw data fit-statistics

	Akhtar et al and generated raw data values (in brackets)			
	R^2	Adjusted R^2	Unadjusted ΔR^2	Adjusted ΔR^2
Model 1	.058 (.057)	.056 (.056)	.058 (.057)	.056 (.056)
Model 2	.203 (.195)	.197 (.190)	.145 (.138)	.141 (.134)
Model 3	.255 (.236)	.245 (.227)	.052 (.041)	.048 (.037)
Model 4	.267 (.295)	.256 (.286)	.012 (.059)	.011 (.059)

Note: all Models differ statistically significantly from one another at p < 0.000001

Although the raw data solution is slightly different in terms of R^2 values, they are close enough to provide sensible comparison.

3.2 Model Comparisons at the level of Engagement observations

There are three ways we can explore the difference between two regressions indexed by their model R^2 values.

3.2.1 Calculate the magnitude agreement between the Predicted observations from two models

The logic here is that if a model is to be viewed as providing predictions which are substantively different from those generated using an alternative model, the agreement between the observations should reflect the lift in accuracy of prediction in a lower valued index of similarity. That is, if the predicted observations from the two models were identical to one another, then the similarity coefficient should appropriately show that identity. If the predicted observations were completely different from one another, then a similarity coefficient should likewise indicate that difference.

Note, we are **not** interested in observation monotonicity as indexed by a correlation coefficient (e.g. Pearson, Gamma, Spearman etc.), but in the absolute agreement between the two predicted values.

A useful coefficient is the Gower[2] index of similarity. Relative to the maximum possible absolute (*unsigned*) discrepancy between the two pairs of observations, the Gower *discrepancy* coefficient indicates the % average absolute discrepancy between all pairs of observations. When expressed as a similarity coefficient *(by subtracting it from 1),* it indicates the % average similarity between all pairs of observations. The similarity coefficient varies between 0 and 1 (or 0% and 100%). So, a Gower similarity coefficient of say 0.90 indicates that relative to the maximum possible absolute (*unsigned*) discrepancy between them, the observations agree *on average* to within 90% of each other's values. Details are provided in Appendix 1.

Table 11: Predicted observation agreement between regression models

	Unadjusted ΔR^2	Adjusted ΔR^2	Gower Agreement
Model 2 vs Model 3	.041	. 037	.98
Model 3 vs Model 4	.059	.059	.97

The impact of adding the seven HPI variables to the five NEO + Age variables increased explained variation by .041 (4.1%). However, the actual impact on the predicted observations is negligible, as the Gower index indicates that the observations predicted by Model 3 agree on average to within 98% of the magnitude of observations from Model 2.

Likewise, the impact of adding trait EI variables to the variables in Model 3 increased explained variation by .059 (5.9%). However, the actual impact on the predicted observations is negligible, as the Gower index indicates that the observations predicted by Model 4 agree on average to within 97% of the magnitude of observations from Model 2.

Conclusion:

From this perspective, there really is 'nothing to see here' using the HPI or Trait EI to predict Engagement scores over and above using the TIPI version of the Big Five.

[2] Gower, J.C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 27, 857-874.

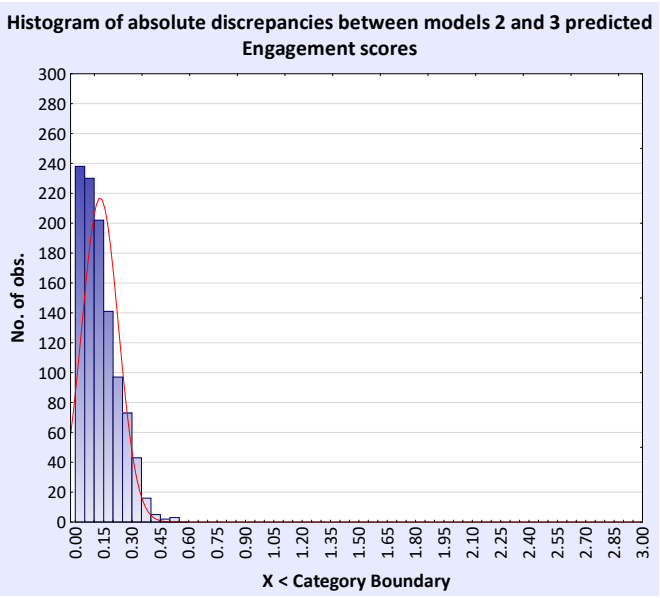
3.2.2 Compute the frequencies of the absolute discrepancies between the predicted values of Engagement from two models

Here the goal is to compute the frequencies of discrepancy magnitudes between the predicted values from each regression model, display them graphically, and express the median discrepancy as a % of the effective measurement range of Engagement [0 to 6].

Table 12: The frequencies of discrepancies between Model 2 and Model 3 predicted values of Engagement

	Frequency table: Model 2 - Model 3: absolute discrepancies			
From To	Count	Cumulative Count	Percent	Cumulative Percent
0.00 <=x<0.05	238	238	22.67	22.67
0.05 <=x<0.10	230	468	21.90	44.57
0.10 <=x<0.15	202	670	19.24	63.81
0.15 <=x<0.20	141	811	13.43	77.24
0.20 <=x<0.25	97	908	9.24	86.48
0.25 <=x<0.30	73	981	6.95	93.43
0.30 <=x<0.35	43	1024	4.10	97.52
0.35 <=x<0.40	16	1040	1.52	99.05
0.40 <=x<0.45	5	1045	0.48	99.52
0.45 <=x<0.50	2	1047	0.19	99.71
0.50 <=x<0.55	3	1050	0.29	100.00
0.55 <=x<0.60	0	1050	0.00	100.00
Missing	0	1050	0.00	100.00

Figure 1: The histogram of discrepancies between Model 2 and Model 3 predicted values of Engagement



The median discrepancy is **0.11**, which given the measurement range for Engagement of [0 to 6] indicates a **1.8%** median discrepancy between Model 2 predicted scores and Model 3 predicted scores.

Conclusion:

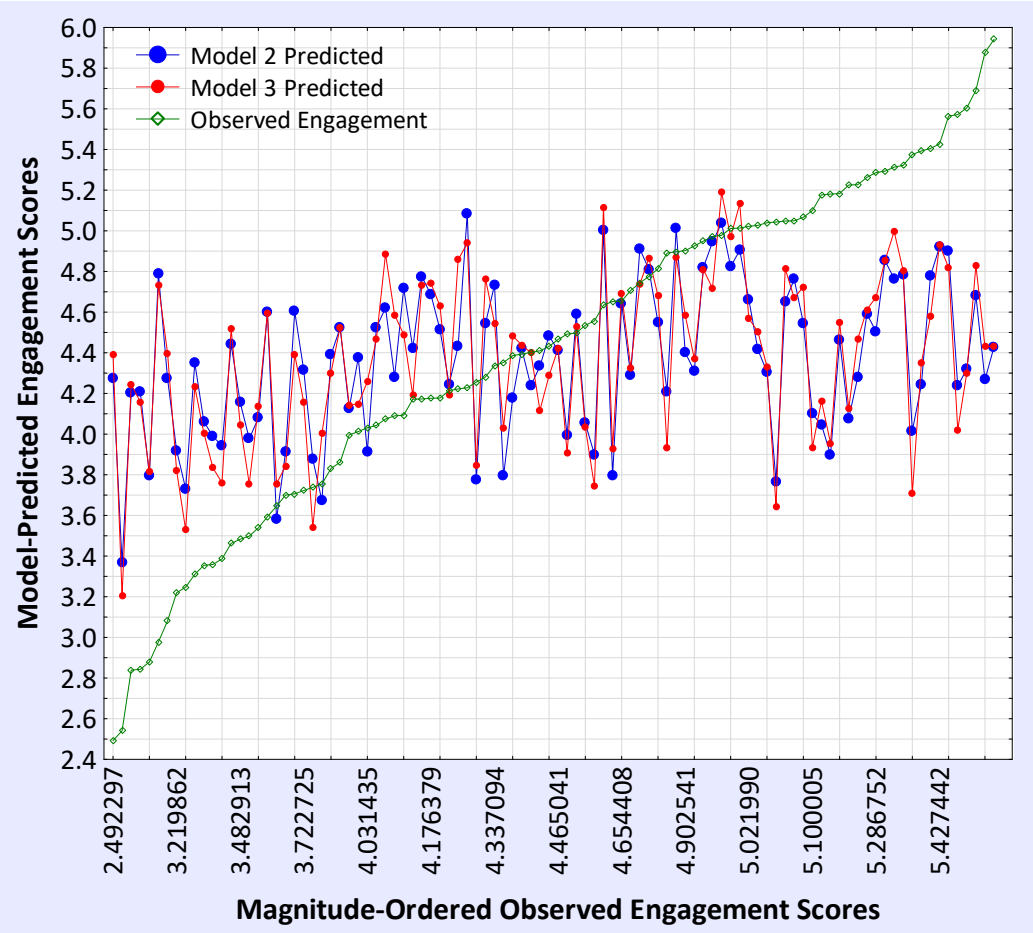
From this perspective, there really is 'nothing to see here' using the HPI to predict Engagement scores over and above using Age and the TIPI version of the Big Five.

3.2.3 Plot the model predicted engagement scores against observed engagement scores

Here, I order the actual Engagement scores and the Models 2 and 3 predicted scores by the observed engagement score magnitude. Then, for visual clarity, subsample 10% of the 1050 cases within the minimum and maximum observed Engagement score range (every 10th observation).

If Model 3 was clearly a better predictor of Engagement, it's observations would lie visibly closer to the observed Engagement scores than Model 2. What we actually see is that neither model predicts Engagement scores accurately, and that both models' predicted values are so similar that there is no meaningful or any sensible interpretable advantage in using Model 3 scores over and above Model 2.

Figure 2: Models 2 and 3 predicted values plotted against Observed Engagement scores



Conclusion:

From this perspective, there really is 'nothing to see here' using the HPI to predict Engagement scores over and above using Age and the TIPI version of the Big Five.

And yes, I haven't bothered looking at the difference between Models 3 and 4 as above because it's obvious a 0.018 increase in ΔR^2 values (between Models 2 vs 3, and Models 3 vs 4) isn't going to produce anything of interest (the Gower is 0.97 between Model 3 and 4 predicted values). There is no "important trait EI" impact.

4. Taking an even closer look at ΔR^2 values

The problem with the Akhtar et al analysis is that the authors chose from the outset to express their observations (except for Age) as min-max threshold-constrained real-values. In fact, they were all integer sum-scores.

Although the re-expression of the scores by dividing the scale sum-scores through by the numbers of items in a scale makes no difference if one is only concerned with presenting summary parameter results *(and relying upon standardized scores for correlations, beta weights etc.),* it makes a huge difference if you wish to evaluate the explanatory accuracy of any model. Why? Because here it is important to gauge how accurately you can predict the observed integer score **(in its own metric)** using your model.

And, when it comes to evaluating the meaningfulness of ΔR^2 values, you need to evaluate whether the discrepancy between observed and predicted scores is actually interpretable in terms of what it means for a score to differ say by 3 integers over a potential ordered-magnitude measurement range of say 60. i.e. what is the behavioural or cognitive difference between an Engagement score of 30 and 33 in a measurement range of 54 (the UWES-9 scale score range)? After all, the scale is no more than a convenience; the assignment of equal-interval integers onto something which we can approximately order from high to low with maybe a few *more or less* meaningful distinctions in-between.

We have no evidence at all that Engagement varies as a quantity *(real-valued, continuous, additive unit metric with a standard unit of measurement; in short, a typical SI physics extensive or derived unit variable)*[3]. So, for an effect to possess some decent pragmatic value, it's going to have to metaphorically stand up and say "look at me". In short we will be able to see it "by eye" – clear as day, when plotted or expressed appropriately as with Observation Oriented Analyses[4] (OOM) analyses or the kind of analyses you've seen above.

So what I want to do here is strip this problem down to its basics. I'm going to generate some new data (as integer sum scores) from a *specified-in-advance* correlation matrix; with just two predictors (Mindfulness and Conscientiousness) and one outcome variable, UWES Engagement scores.

Table 13: A hypothetical correlation matrix between Engagement, Mindfulness, and Conscientiousness

	Initial hypothetical correlation matrix		
	1 Engagement	2 Mindfulness	3 Conscientiousness
Engagement	1	0.3	0.13
Mindfulness	0.3	1	0.05
Conscientiousness	0.13	0.05	1
Means	27	25	10
Std.Dev.	6	5	3

The scale score ranges are:
Engagement [0-54, as UWES]
Mindfulness [0-40]
Conscientiousness [0-20]

Then I generated 1000 cases of normally distributed real-valued data using the same method as before, with constraints in place as per the scale score minimum and maximum values. These 'scores' were then re-expressed

[3] http://physics.nist.gov/cuu/Units/units.html
[4] Grice, J. (2014). Observation Oriented Modeling: Preparing students for research in the 21st century. *Innovative Teaching*, 3, 1-27
Grice, J. (2015). From means and variances to persons and patterns. *Frontiers in Psychology: Quantitative Psychology and Measurement* (http://dx.doi.org/10.3389/fpsyg.2015.01007), 6:1007, 1-12.

as rounded integers, as would be used when scale score are expressed as sum-scores. Correlating these integers produced the correlation matrix in Table 14.

Table 14: The hypothetical correlation matrix (from integer scores) between Engagement, Mindfulness, and Conscientiousness

Variable	Pearson correlations, n=1000 cases Using generated data - integer scores		
	Engagement	Mindfulness	Conscientiousness
Engagement	1.00	0.29	0.12
Mindfulness	0.29	1.00	0.04
Conscientiousness	0.12	0.04	1.00

Basically- good enough for purpose.

Table 15: The descriptive statistics of the three integer variables

Variable	Descriptive Statistics for three variable problem					
	Valid N	Mean	Median	Minimum	Maximum	Std.Dev.
Engagement	1000	27.0	27	12	42	5.786
Mindfulness	1000	25.0	25	12	38	4.817
Conscientiousness	1000	10.0	10	2	18	2.903

4.1. The regression models

Model 1: Engagement predicted by Mindfulness
Model 2: Engagement predicted by Mindfulness and Conscientiousness

Table 16: The three variable problem: Model R^2 and ΔR^2

	Unadjusted R^2	**Adjusted R^2**	**Unadjusted ΔR^2**	**Adjusted ΔR^2**
Model 1	.0864	.0855	.0864	.0855
Model 2	.0977	.0959	.0113	.0104

Note the ΔR^2 value of 0.011 for Model 2 is only 0.001 different from the ΔR^2 value of 0.012 for Model 4 in Akhtar et al, for the incremental 'effect' of Trait EI.

So, let's run through our evaluation analyses for the 0.011 ΔR^2

4.1.1 Calculate the magnitude agreement between the Predicted observations from the two models

The Gower index is: **0.99.** Which indicates that relative to the maximum possible absolute (*unsigned*) discrepancy between them, the observations agree *on average* to within 99% of each other's values.

4.1.2 Compute the frequencies of the absolute discrepancies between the predicted values of Engagement from the two models

Here the goal is to compute the frequencies of absolute magnitudes of discrepancy between the predicted values from each regression model, display them graphically, and express the median discrepancy as a % of the effective measurement range of Engagement [0 to 54].

Table 17: The frequencies of absolute *real-valued* prediction discrepancies between Model 1 and Model 2 predicted values of Engagement

	Frequency table: Model 1 - Model 2: absolute discrepancies			
From To	Count	Cumulative Count	Percent	Cumulative Percent
0.000 <=x<0.100	134	134	13.40	13.40
0.100 <=x<0.200	82	216	8.20	21.60
0.200 <=x<0.300	172	388	17.20	38.80
0.300 <=x<0.400	53	441	5.30	44.10
0.400 <=x<0.500	160	601	16.00	60.10
0.500 <=x<0.600	17	618	1.70	61.80
0.600 <=x<0.700	146	764	14.60	76.40
0.700 <=x<0.800	1	765	0.10	76.50
0.800 <=x<0.900	107	872	10.70	87.20
0.900 <=x<1.000	3	875	0.30	87.50
1.000 <=x<1.100	59	934	5.90	93.40
1.100 <=x<1.200	8	942	0.80	94.20
1.200 <=x<1.300	30	972	3.00	97.20
1.300 <=x<1.400	8	980	0.80	98.00
1.400 <=x<1.500	11	991	1.10	99.10
1.500 <=x<1.600	6	997	0.60	99.70
1.600 <=x<1.700	3	1000	0.30	100.00
1.700 <=x<1.800	0	1000	0.00	100.00
Missing	0	1000	0.00	100.00

87.5% of predicted observations from Model 2 differ by less than ±1 integer magnitude from Model 1's predicted values.

The median discrepancy is 0.42, which when expressed relative to the possible measurement range of the Engagement scores, is a **0.8%** 'lift' in accuracy. That says it all.

Still not convinced? Then let's convert the predicted values (which are real-valued) into rounded integers, to match the metric of the Engagement scores. If we now compute the absolute discrepancy between Model 1 and Model 2 predicted Engagement values and compute the frequency distribution of these prediction discrepancies, we see:

Table 18: The frequencies of absolute integer-prediction discrepancies between Model 1 and Model 2 predicted values of Engagement

	Frequency table: Model1 - Model 2 integers: absolute values			
Category	Count	Cumulative Count	Percent	Cumulative Percent
0	508	508	50.80	50.80
1	466	974	46.60	97.40
2	26	1000	2.60	100.00
Missing	0	1000	0.00	100.00

Yep, you read it right … 97.4% of predicted observations from Model 2 differ by equal to or less than ±1 integer magnitude from Model 1's predicted values.

Conclusion:

Again, this final analysis shows there really is 'nothing to see here'.

I'm not even going to bother with the graph of predicted values against Engagement, we already know from the above that a 0.011 ΔR^2 is simply irrelevant to any claim of incremental effect'.

But we can at least enjoy the desolation depicted in the image on the left! A visual metaphor if you like of the intellectual and scientific sterility of trivial ΔR^2 values.

Akhtar et al, with their 0.012 ΔR^2 value for Model 4 incremental Trait EI 'effect' have presented a result with absolutely no theoretical or practical consequences at all.

And, as is typical of all those who present these ridiculously trivial incremental effect sizes, they nevertheless go on to suggest substantive actions on the behalf of practitioners/HR:

> "The results of this study have both theoretical and applied implications. On a practical level by understanding dispositional predictors of engagement, organizations can select employees high on the personality traits examined in this study, specifically EI, openness, extraversion, conscientiousness, adjustment, ambition, and interpersonal sensitivity. By including these personality characteristics in their selection criteria, organizations can improve the likelihood of finding high-performing job candidates that other selection systems may exclude." p. 48, under the heading "Implications".

I'm sorry, but this is bad advice given the results they have presented. Yet, they and many other students/researchers do exactly the same, and wonder why psychologists become sources of ridicule and invite indifference from others, rather than respect. This is one of the issues addressed by Chris Ferguson in his article: Ferguson, C.J. (2015). "Everybody knows psychology is not a real science": Public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community, and the general public. *American Psychologist*, 70, 6, 527-542.

Abstract

In a recent seminal article, Lilienfeld (2012) argued that psychological science is experiencing a public perception problem that has been caused by both public misconceptions about psychology, as well as the psychological science community's failure to distinguish itself from pop psychology and questionable therapeutic practices. Lilienfeld's analysis is an important and cogent synopsis of external problems that have limited psychological science's penetration into public knowledge. The current article expands upon this by examining internal problems, or problems within psychological science that have potentially limited its impact with policymakers, other scientists, and the public. These problems range from the replication crisis and defensive reactions to it, overuse of politicized policy statements by professional advocacy groups such as the American Psychological Association (APA), and continued overreliance on mechanistic models of human behavior. It is concluded that considerable problems arise from psychological science's tendency to over-communicate mechanistic concepts based on weak and often unreplicated (or unreplicable) data that do not resonate with the everyday experiences of the general public or the rigor of other scholarly fields. It is argued that a way forward can be seen by, on one hand, improving the rigor and transparency of psychological science, and making theoretical innovations that better acknowledge the complexities of the human experience.

5. So what value of ΔR^2 should be taken seriously?

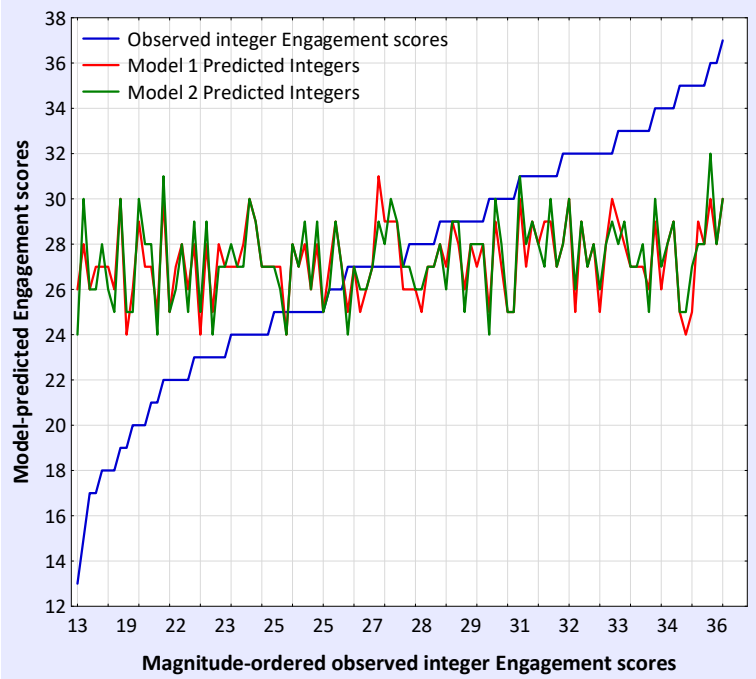
This is tricky, because it depends on the hypothesis being proposed. If, for example, we proposed that Trait EI matters substantively to a person's reported Engagement, then observing a tiny ΔR^2 is important because it is clear empirical evidence against such a proposition. So, it truly matters from a theoretical perspective.

However, most researchers are looking for a substantive finding, because this will almost certainly have important practical utility. So, observing the tiny ΔR^2 is again important because it means there is no 'effect' with any pragmatic or practical value.

Small effects can be important when considering epidemiological *(population-based*) effects which have a truly important outcome *(e.g. socio-economic status, physical health, disease/virus-incidence, personal mortality)*. But the phenomena being predicted in I/O psychology studies are not 'population-relevant' except in broad employment/workplace-relevant descriptive terms *(such as the pervasive influence of being "Conscientious")*.

Given a multiple *R* of 0.31 for our two-variable prediction of Engagement, some might be tempted to conclude that this must surely translate into a positive benefit for an organization if the function is used as a selection screen for candidates. But look at the plot of the predicted values against actual Engagement scores, where the x-axis Engagement score magnitudes are ordered from low to high observed scores *(using a 10% subsampling across the range of observed Engagement scores, for plot-clarity)*:

Figure 3: Models 1 and 2 predicted Engagement values plotted against magnitude-ordered Observed Engagement scores



The predictions are clearly only accurate over a small 'middle-range' of Engagement scores. To see this another way, let's say that everyone who scores just above average (28) on Engagement is classed as a 'success'. So, we now look for the optimal cut-score on Model 1 and Model 2 predicted values, to establish the pragmatic accuracy of using either model as a selection pre-screen for *'likely-to-be-Engaged'* employees. Remember, as per Akhtar et al, we have just recommended that the extra variable in the equation should be assessed in a prediction model *(here it is Conscientiousness; in their model it was Trait EI, where both are relying upon a similar ΔR^2 value of .011 or .012)*. A cut-score optimisation analysis for Model 1 reveals:

Figure 4: Cut-score optimisation for Model 1 predicted Engagement

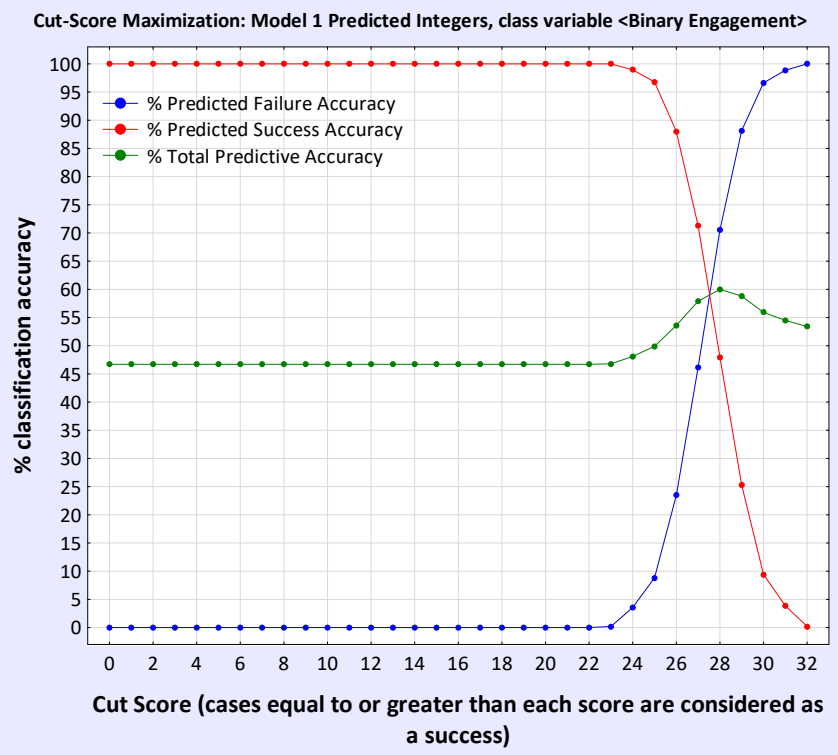
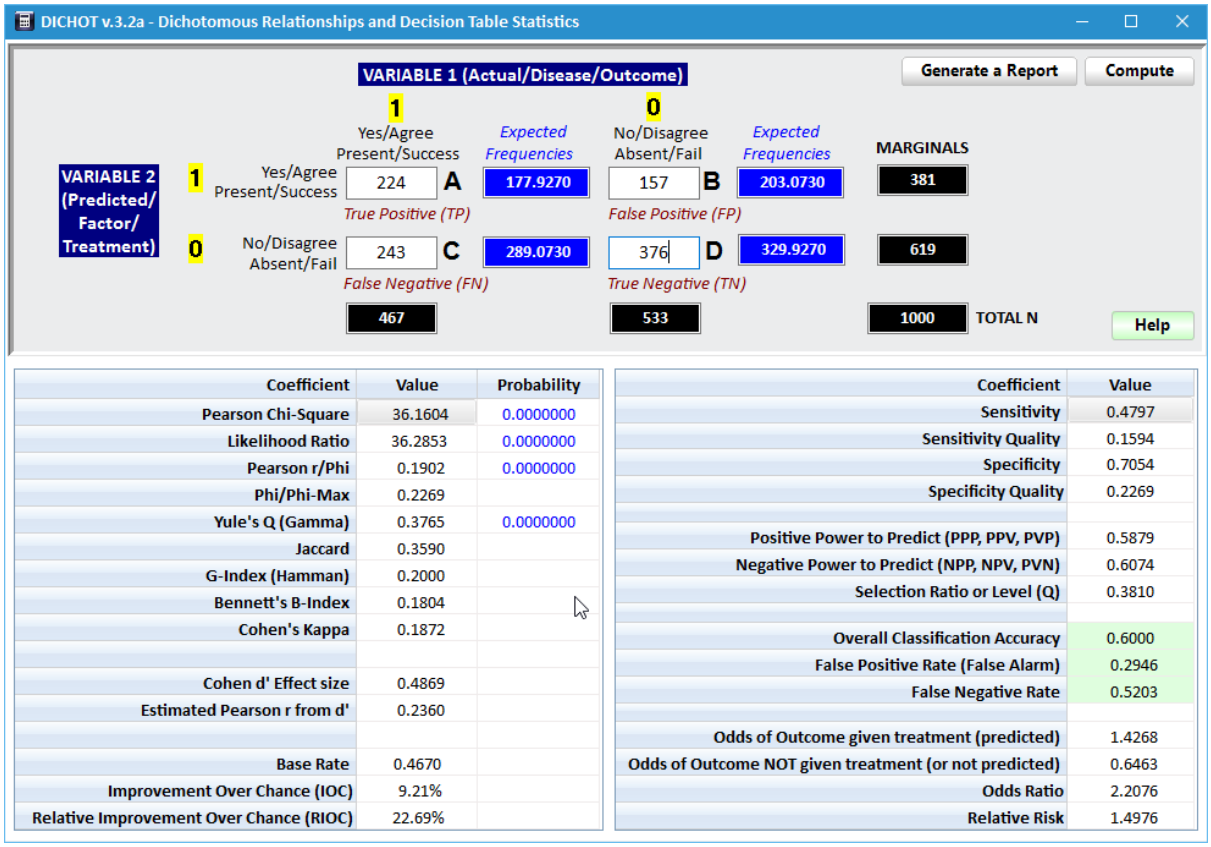


Figure 5: Comprehensive actuarial analysis for Model 1 optimal cut-score of 28



Coefficient	Value	Probability
Pearson Chi-Square	36.1604	0.0000000
Likelihood Ratio	36.2853	0.0000000
Pearson r/Phi	0.1902	0.0000000
Phi/Phi-Max	0.2269	
Yule's Q (Gamma)	0.3765	0.0000000
Jaccard	0.3590	
G-Index (Hamman)	0.2000	
Bennett's B-Index	0.1804	
Cohen's Kappa	0.1872	
Cohen d' Effect size	0.4869	
Estimated Pearson r from d'	0.2360	
Base Rate	0.4670	
Improvement Over Chance (IOC)	9.21%	
Relative Improvement Over Chance (RIOC)	22.69%	

Coefficient	Value
Sensitivity	0.4797
Sensitivity Quality	0.1594
Specificity	0.7054
Specificity Quality	0.2269
Positive Power to Predict (PPP, PPV, PVP)	0.5879
Negative Power to Predict (NPP, NPV, PVN)	0.6074
Selection Ratio or Level (Q)	0.3810
Overall Classification Accuracy	0.6000
False Positive Rate (False Alarm)	0.2946
False Negative Rate	0.5203
Odds of Outcome given treatment (predicted)	1.4268
Odds of Outcome NOT given treatment (or not predicted)	0.6463
Odds Ratio	2.2076
Relative Risk	1.4976

For Model 1 we have 60% overall classification accuracy, with a 29% False-positive rate and 52% False-negative rate. Our overall error-count is 400 cases from a 1000-case sample.

Figure 6: Cut-score optimisation for Model 2 predicted Engagement

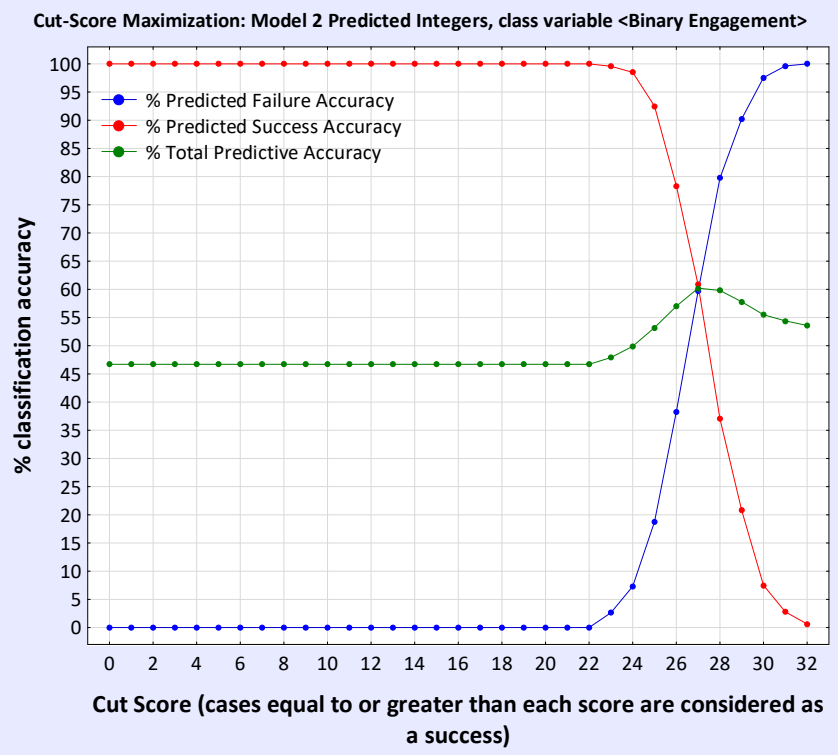
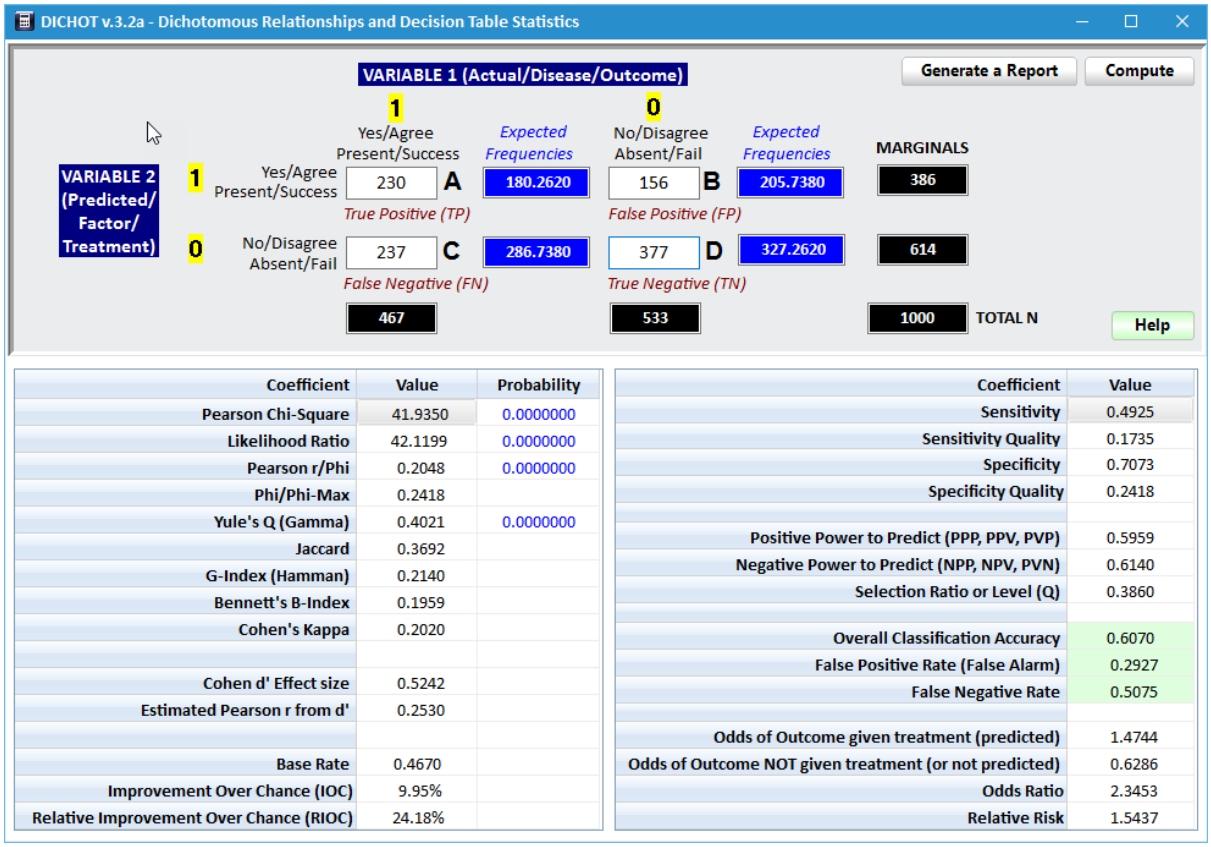


Figure 7: Comprehensive actuarial analysis for Model 2 optimal cut-score of 28



Coefficient	Value	Probability
Pearson Chi-Square	41.9350	0.0000000
Likelihood Ratio	42.1199	0.0000000
Pearson r/Phi	0.2048	0.0000000
Phi/Phi-Max	0.2418	
Yule's Q (Gamma)	0.4021	0.0000000
Jaccard	0.3692	
G-Index (Hamman)	0.2140	
Bennett's B-Index	0.1959	
Cohen's Kappa	0.2020	
Cohen d' Effect size	0.5242	
Estimated Pearson r from d'	0.2530	
Base Rate	0.4670	
Improvement Over Chance (IOC)	9.95%	
Relative Improvement Over Chance (RIOC)	24.18%	

Coefficient	Value
Sensitivity	0.4925
Sensitivity Quality	0.1735
Specificity	0.7073
Specificity Quality	0.2418
Positive Power to Predict (PPP, PPV, PVP)	0.5959
Negative Power to Predict (NPP, NPV, PVN)	0.6140
Selection Ratio or Level (Q)	0.3860
Overall Classification Accuracy	0.6070
False Positive Rate (False Alarm)	0.2927
False Negative Rate	0.5075
Odds of Outcome given treatment (predicted)	1.4744
Odds of Outcome NOT given treatment (or not predicted)	0.6286
Odds Ratio	2.3453
Relative Risk	1.5437

For Model 2 we have 60.7% overall classification accuracy, with a 29% False-positive rate and 51% False-negative rate. Our overall error-count is 393 cases from a 1000-case sample.

Our 'incremental effect' has translated into a 0.7% increase in predictive accuracy. That is what Akhtar et al were basing their recommendation on, for the inclusion of Trait EI into a selection battery of assessments.

And still I hear some responding 'well, better 0.7% than 0%'. The problem with that line of argument is that it presupposes above-average 'Engagement' relates perfectly to some important outcome. It doesn't. We are probably looking at a 0.3 correlation *at best* between Engagement and any recognizable organizational outcome. Expressed as a binomial effect size display (BESD)[5] accuracy estimate, it translates to about 65% classification accuracy, a lift of about 15% above chance, so our .7% is probably nearer .7 x .65 = .**45%.**

In short, it's a hopeless case trying to justify small ΔR^2 values unless very special conditions hold, requiring careful, empirical cost-benefit analysis involving the evaluation of the implications of failures and successes.

So, in answer to my two questions:

❶ Does a small ΔR^2 value have any pragmatic value at all?

Of course not. The analyses above show that small values such as those from my simulations or the published results from Akhtar et al quite literally useless for all practical and theoretical purposes, except in the latter case as evidence AGAINST any theory claim which states an incremental effect should have been substantive.

❷ What magnitude of ΔR^2 is worth reporting as more than "*nothing to see here*"?

As I say, tricky … the analyses I undertook show just how far you have to dig into your observations in order to figure out the pragmatic utility of a ΔR^2 value.

ΔR^2 values have to be evaluated for their consequences in the actual metric of the observed variable being 'accounted for'; not its standardized form, 'latent' form, or some other 'transformed' version of its observations.

[5] http://www.pbarrett.net/techpapers/BESD_April_2013.pdf

Appendix 1: The Gower Agreement Coefficient

Relative to the maximum possible absolute (*unsigned*) discrepancy between the two pairs of observations, the Gower *discrepancy* coefficient indicates the % average absolute discrepancy between all pairs of observations. When expressed as a similarity coefficient (by subtracting it from 1), it indicates the % average similarity between all pairs of observations. The Gower coefficient varies between 0 and 1 (or 0% and 100%).

So, a Gower similarity coefficient of say 0.90 indicates that relative to the maximum possible absolute (*unsigned*) discrepancy between them, the observations agree on average to within 90% of each other's values.

If you change the value of that maximum possible discrepancy, then the Gower coefficient will change to reflect this, as the discrepancies between pairs of observations are divided (scaled) by that maximum possible discrepancy value. E.g. if two observations differ by 5, and the measurement range of each observation is 10, then the relative discrepancy is 0.5. However, if the measurement range for each observation was say 100, then the relative discrepancy would be just 0.1.

But that's the whole point of the Gower, it tells you how discrepant (or similar) observations are, RELATIVE to how maximally discrepant they could have been.
A 5-point difference in a 10-point maximum measurement range is substantial.
A 5-point difference between observations within a 100-point measurement range is trivial.

The equation for the Gower similarity index is:

$$Gower_{similarity} = 1 - \left[\frac{\sum_{i=1}^{n} \left(\frac{|obs_{1i} - obs_{2i}|}{range} \right)}{n} \right]$$

$n =$ the number of cases

$range =$ the maximum possible discrepancy between the two attribute/variable magnitudes

$obs_{1i} =$ the observed value for case i of n for dataset or attribute 1

$obs_{2i} =$ the observed value for case i of n for dataset or attribute 2

A free-to-download computer program for computing the Gower, along with a free bootstrap program to compute its statistical significance *(in terms of the likelihood of observing a coefficient as large as computed by chance alone)* are available from:

http://www.pbarrett.net/Gower/Gower.html and http://www.pbarrett.net/Bootstrap/Bootstrap.html