Looking for Validity or Testing It?

The Perils of Stepwise Regression, Extreme-Scores Analysis, Heteroscedasticity, and Measurement Error

John Antonakis
john.antonakis@unil.ch
Department of Organizational Behavior
Faculty of Business and Economics
University of Lausanne
Internef 618
1015 Lausanne, Switzerland
Tel.: +41 21 692 3438

Joerg Dietz
joerg.dietz@unil.ch
Department of Organizational Behavior
Faculty of Business and Economics
University of Lausanne
Internef 617
1015 Lausanne, Switzerland
Tel.: +41 21 692 3682

IN PRESS

Personality and Individual Differences

Looking for Validity or Testing It?

The Perils of Stepwise Regression, Extreme-Scores Analysis, Heteroscedasticity, and Measurement Error

When researchers introduce a new test they have to demonstrate that it is valid, using unbiased designs and suitable statistical procedures. In this article we use Monte Carlo analyses to highlight how incorrect statistical procedures (i.e., stepwise regression, extreme scores analyses) or ignoring regression assumptions (e.g., heteroscedasticity) contribute to wrong validity estimates. Beyond these demonstrations, and as an example, we re-examined the results reported by Warwick, Nettelbeck, and Ward (2010) concerning the validity of the Ability Emotional Intelligence Measure (AEIM). Warwick et al. used the wrong statistical procedures to conclude that the AEIM was incrementally valid beyond intelligence and personality traits in predicting various outcomes. In our re-analysis, we found that the reliability-corrected multiple correlation of their measures with personality and intelligence was up to .69. Using robust statistical procedures and appropriate controls, we also found that the AEIM did not predict incremental variance in GPA, stress, loneliness, or well-being, demonstrating the importance for testing validity instead of looking for it.

Keywords: emotional intelligence; general intelligence; personality; validity; errors-in-variables; heteroscedasticity; truncation; Monte Carlo.

Introduction

Tavris and Aronson (2007, p. 108) noted that "the scientific method consists of the use of procedures designed to show not that our predictions and hypothesis are right, but that they might be wrong." This statement is germane to validity testing of new measures and it is important for science to guard against the proliferation of tests that might not explain something new. Scientists should use fair procedures that allow not only for establishing but also for falsifying the validity of their measures. Yet at times, scientists employ weak statistical procedures that may maximize the likelihood of finding validity.

The litmus test for determining the utility of a new measure is incremental validity. Meehl (1959, p. 125) referred to this test as the "most pressing *immediate* clinical research problem." A new test must show some "*increment* in predictive efficiency" beyond established measures (Sechrest, 1963, p. 154), and this increment will only emerge if the new test taps unique variance not predicted by current measures. Establishing incremental validity is a theory-driven exercise wherein the researcher reviews past research on a criterion to identify established predictors that will be used as control variables (Sechrest, 1963). In terms of including controls, which are usually entered first in the regression, it is advisable to be conservative and to include more rather than fewer variables to avoid "omitted variable bias" (Cameron & Trivedi, 2005).

Omitting variables from a regression model biases the coefficients of the remaining variables to the extent that the omitted variables correlate with other variables in the model (Cameron & Trivedi, 2005). Thus, important theoretical control variables should never be dropped from regression models: Even if not significant individually; they might be jointly significant and these multivariate effects are necessary for adequate regression adjustment. In the second step, the new measure (or measures) is added and the test of the significance of its coefficient (or coefficients) or a nested *F*-test will show whether the *r*-square changes significantly.

Theory-driven hierarchical approaches are always preferable to data-driven methods like stepwise regression, wherein an algorithm "selects" the variables that will be entered (Copas, 1983; Leigh, 1988; Thompson, 1995). Beyond the regression approach used, it is imperative that the assumptions of the regression estimator are met. These assumptions include homoscedastic regression residuals, perfectly-measured independent variables, and a non-truncated sample (Cameron & Trivedi, 2005; Draper & Smith, 1998). If any of these assumptions are violated, estimate consistency (i.e., accuracy) may be affected as would inference (i.e., standard errors of estimates).

The purpose of our research is to demonstrate the perilous effects of using incorrect procedures in validity testing. Employing Monte Carlo simulations, we provide simple and visual evidence to show the consequences of using wrong methods. We reanalyzed the data of Warwick et al., (2010)—which we requested and obtained for verification and reanalysis— who suggested that their new measure of emotional intelligence demonstrated incremental validity. Their findings, though, resulted from using biased estimation procedures and violating assumptions of regression analysis. Using fair and statistically robust procedures, our results suggest that their measures should not be used because of their poor concurrent, divergent, and incremental validity.

On Aiming the Wrong Way: The AEIM Validation Study

Warwick et al. (2010) concluded that their Ability Emotional Intelligence Measure (AEIM) incrementally predicted GPA and other outcomes beyond general cognitive ability and personality. They proposed two kinds of emotional intelligence (EI) scores: *consensus* scores, which benchmark individual responses against the most frequently-endorsed responses, and *confidence* scores, which are self-report measures of confidence in one's responses. Their findings might be potentially important in light of the debate both regarding the construct of EI and the validity of EI tests (Antonakis, Ashkanasy, & Dasborough, 2009; Antonakis & Dietz, 2010; Davies, Stankov, & Roberts, 1998; Locke, 2005; Zeidner, Roberts, & Matthews, 2008). Advocates and adversaries of EI hotly contest the

incremental validity of EI for performance outcomes above and beyond cognitive intelligence and personality traits. The methods used by Warwick et al. (2010), however, are flawed in four respects:

1. Warwick et al. (2010) said they used “hierarchical regressions” yet they noted immediately afterwards for their first tests that their “Stepwise independent variables [were] cognitive ability, significant personality variables, and consensus and confidence scores” (p. 69). It is unclear which estimator or exploratory algorithm they used or which control variables they included.

2. Warwick et al. (2010) formed four subgroups as a function of low and high consensus and confidence AEIM scores. Within these groups, they kept the top and bottom 35% of the sample on these two variables (thus retaining 50.37% of the total observations or 137 of 272 participants). For each subgroup (between 22 and 41 participants), Warwick et al. repeated their regression analyses, arguing “that there is currently considerable variation in education about emotion knowledge [and] therefore likely . . . notable differences in test scores,” which “might be masked by consideration of average scores alone” (p. 67).

3. For tests of the models that predicted outcome variable we found that the residuals were heteroscedastic (Breusch & Pagan, 1979); however, Warwick et al. (2010), neither reported nor attended to the heteroscedasticity problem.

4. Warwick et al. (2010) did not correct for imperfectly measured regressors (i.e., variables that are not perfectly reliable), which included all independent variables (EI, personality, and cognitive ability).

Although the general problems with these procedures—that is, stepwise regression (or failing to include all control variables), extreme scores analysis, heteroscedasticity, and measurement error—have been pointed out in the methodological literature (Bollen, 1989; Cameron & Trivedi, 2005; Copas, 1983; Maxwell & Delaney, 1993; White, 1980), these problems do not seem well understood in applied psychology research. We clarify the nature of these problems in detail below.

Sidestepping Validity with Stepwise Regression

Stepwise regression is a procedure that is not consistent with a theory-driven approach to testing incremental validity. Its exploratory algorithm capitalizes on chance to select predictor variables and ignores the significance of sets of control variables. It produces wrong *F*-tests, biased *R*-squares, and wrong *p*-values because the assumptions of these statistics are violated by how stepwise procedure conducts the "tests" (Copas, 1983; Leigh, 1988; Thompson, 1995). In Warwick et al.'s (2010) data (see Table 1), all controls correlated to an extent with EI and outcomes; thus they should have been retained in all regression models (particularly for their possible multivariate significance).

Using their full sample, Warwick et al. (2010) reported that: "Consensus outcomes significantly negatively predicted loneliness (β = -.13, t = -2.15, p < .05) after controlling for cognitive ability and personality (F(5,266) = 26.30, p < .001) accounting for 2% additional variance" (p. 69). We could not reproduce their findings, using the same set of predictors that they apparently used (i.e., the AEIM consensus and confidence scores, cognitive ability, and the two personality factors that correlated significantly with loneliness at p < .05, that is extraversion and neuroticism) and a normal variance estimator that was not robust to violations of heteroscedasticity (as they apparently did). Our results showed that the AEIM consensus measure did not significantly predict loneliness despite the sufficiently large sample size of 272 to detect significant effects: β = -.10, t = -1.75, p > .05 (regression F(5, 266) = 35.99, p < .001).

[Insert Table 1 here]

That Warwick et al. reported a significant β (at p < .05) for consensus scores is puzzling. Therefore, we also tried various combinations of variables and specifications with stepwise regression to reproduce what they did but we could not obtain a β of -.13 for the consensus score. If Warwick et al. used a stepwise algorithm, they used probability cut-offs for entering and removing variables from the model that we as independent researchers could not possibly guess and reproduce. Alternately,

Warwick et al. might have incorrectly reported their result or maybe did not use stepwise regression, a possibility we consider.

Because stepwise regression is still used by many researchers, we demonstrate explicitly how stepwise regression can produce specious findings from simple random variables. We conducted a Monte Carlo simulation (20 replications), varying sample sizes from $n = 20$ to $n = 45$ (the approximate sample sizes in Warwick et al.'s (2010) analyses of "extreme scores"). We also varied the number of predictors included in the model from two to nine, nine being the number of predictors in Warwick et al.'s study. Note that the predictors and the dependent variable (y) were random variables drawn from a normal distribution with a mean of zero and standard deviation of 1. We also included a random heteroscedastic error term in predicting y, conditional on one of the covariates. We then used stepwise regression with a normal variance estimator to predict y, such that our stepwise simulations were done with backward selection at a significance level for removal of .20. As Figure 1 shows, the stepwise regression model miraculously "found" combinations of significant variables and predicted r-squares varying between .14 and .37 (just from pure noise)! And as Figure 2 shows, the F-tests of the models approached significance even with only five variables in the model! If Warwick et al. (2010) used stepwise regression, our simulations suggest that their results are dubious.

[Insert Figures 1 and 2 here]

Divide and Conquer: Extreme Scores (Subgroup) Analyses

In testing for incremental validity, the benefits of using the entire sample are obvious. It is a precondition for interpreting population estimates, and it facilitates comparisons of validity coefficients across different samples. Warwick et al. (2010) created artificial subgroups and analyzed extreme because apparently average scores attenuate relations. Such procedures, though, are severely flawed and the deletion of parts of a sample can severely bias estimates. It might be defensible to remove clear outliers (e.g., 3 *SDs* from the mean); however, to delete 50% of the sample in the middle of the

distribution as Warwick et al. did has no justification other than artificially benefitting from using extreme scores. Deleting these middle values results in an obvious misrepresentation of the data (see note in Figure 3). Furthermore, chopping up samples into selected groups across two independent variables leads to false findings (Maxwell & Delaney, 1993). Nobel prizes have been earned in econometrics for methods to correct for truncated samples, among other contributions (e.g., Heckman, 1979; Tobin, 1958). In short, researchers must avoid or correct sample bias instead of creating it.

We conducted another Monte Carlo analysis to show that relations between variables are actually attenuated in *any* truncated group of the sample (i.e., both in the extremes and the middle). Suppose that in a population of individuals (n = 272 as in Warwick et al.'s study), the true model that generated the relation between x and y is:

$$x = 5 + r \quad \text{(eq. 1)}$$

$$y = 2 + 1*x + 2*e \quad \text{(eq. 2)}$$

where r is an independent random variable from a normal distribution with a mean of 1 and a standard deviation of 0 and where e is an independent random error term with a mean of 1 and a standard deviation of 0. As Figure 3 shows, when splitting the sample into three groups on x (on high, middle, and low values of x, using the top and bottom 35%, as Warwick et al. did), the relation between x and y is attenuated in *all* groups, not just in the average scores on x.

[Insert Figure 3 here]

Estimating models in subgroups can not only attenuate relations. It can produce truly inaccurate findings as the next Monte Carlo simulation shows, particularly in conjunction with multiple correlated covariates in the model. We simulated data where y depends on nine independent variables (the number of independent variables in Warwick et al.'s study), some of which are correlated with each other. The model that generated the data was the following (n = 272):

$$x1 = r1 \quad \text{(eq. 3)}$$

$x2 = r2 + .2*x1$ (eq. 4)

$x3 = r3 + .3*x1 + .3*x2$ (eq. 5)

$x4 = r4 + .2*x1$ (eq. 6)

$x5 = r5$ (eq. 7)

$x6 = r6 + .2*x5$ (eq. 8)

$x7 = r7 + .5*x5$ (eq. 9)

$x8 = r8 + .1*x7$ (eq. 10)

$x9 = r9 + .1*x7$ (eq. 11)

$y = 5 + 1*x1 + 1*x2 + 1*x3 + 1*x4 + 1*x5 + 1*x6 + 1*x7 + 1*x8 + 1*x9 + 2*e$ (eq. 12)

where *r*1-*r*9 are independent random variables (with mean 0 and *SD* 1) and *e* is an independent random term conditioned on *x*1 (to make *e* heteroscedastic). Thus, the true model parameters (estimated mean coefficients) for the nine *x*'s should be 1. A Monte Carlo simulation using standard regression that included all covariates and with a robust variance estimator recovered the true estimates almost precisely. Stepwise regression in subgroupings (and with a normal variance estimator), in contrast, was not able to do so (see Figure 4). It consistently selected the wrong number of regressors, had incorrect *r*-squares, and biased estimates. In one of the subgroupings (i.e., in the High-Low condition), the mean beta coefficient of *x1* was 339% higher than the true value. It is important to note how inflated the coefficients were *for the two variables on which the cut-offs were conditioned!*

[Insert Figure 4 here]

These Monte Carlo simulations suggest that Warwick et al.'s (2010) truncation of their data and the subsequent use of stepwise regression in subgroups likely produced erroneously high coefficients for these two EI scores. *Even with regular regression analysis* (supposing that Warwick et al. did not use stepwise regression), a Monte Carlo simulation showed that *x*1, for example, had a coefficient of 1.79 in the "high-high" group (i.e., 79% higher than it should be). We trust that it is now evident that

regression analyses within extreme-score groups, using normal or stepwise regression, leaves much to be desired as an approach to psychometric validity testing.

Being Heteroclitic with Heteroscedasticity

Residuals of regression models must not be heteroscedastic. Although coefficients are estimated consistently, heteroscedastic residuals result in incorrect estimates of the variance, leading to uninterpretable *t* statistics for the parameter and thus wrong *F* tests for the regression model (White, 1980). Using robust variance estimators or bootstrapping of standard errors is necessary to ensure inference consistency (Ando & Hodoshima, 2007). Given that the problems of heteroscedasticity are largely ignored, we used Monte Carlo analysis once more using the following generated data:

$$x1 = 5*r1 \quad \text{(eq. 13)}$$

$$x2 = .2*x1 + 5*r2 \quad \text{(eq. 14)}$$

$$y = 1+ .45*x1 + .70*x2 + e \quad \text{(eq. 15)}$$

where $r1$ and $r2$ are independent random variables and e is a random term conditioned on $x1$ (to make e heteroscedastic). We ran two Monte Carlo's (k = 20; n = 272): One using regression and a robust variance estimator and one with regression and a normal variance estimator. The mean *t*-statistics for $x1$ and $x2$ for the robust estimator were 1.19 and 2.06 respectively; however, for the normal variance estimator they were 2.07 and 1.79 respectively, leading to incorrect (and opposite) inference!

Erring when Measuring

In addition to the abovementioned limitations, Warwick et al. (2010) ignored measurement error (i.e., their variables were latent). This violation of yet another assumption of model estimation leads to inconsistent estimates (Bollen, 1989). These estimates will not converge to the true population values even with an increasing sample size and the measurement-error bias will transmitted to other variables that correlate with the problematically-measured variables (Cameron & Trivedi, 2005). Methods such as errors-in-variables regression exist to correct for this bias using least-squares or

maximum-likelihood estimation (Draper & Smith, 1998). Because the problem of unreliable measurement is well known in psychological research (Ree & Carretta, 2006; Schmidt & Hunter, 1999), and because of space limitations, we do not provide Monte Carlo results to show this bias.

Having discussed why the results of Warwick et al. (2010) might not theoretically be correct, we reexamined the validity of the AEIM using the correct statistical techniques as described below.

On Being On the Spot

In our reanalysis of the Warwick et al. data, we used errors-in-variables regression (Draper & Smith, 1998), accounting for measurement error with the reliabilities of the scales as is done with maximum likelihood estimation (see Bollen, 1989). Because of the heteroscedasticity problem we bootstrapped standard errors using 1,000 replications (Ando & Hodoshima, 2007). Given that Warwick et al. included an overall cognitive ability score we used this measure and separately its two sub-components, fluid and verbal intelligence.

[Insert Table 2 here]

In the initial analysis, we regressed the AEIM scores on intelligence and personality, predicting a large amount of variance in the AEIM consensus scores (see Table 2). For example, the cognitive ability score had a partial (standardized) coefficient of .69 in predicting the consensus measure! This result suggests that consensus score may be largely redundant (due to low discriminant validity) as a predictor of performance or other outcomes. Note, the zero-order correlation of cognitive ability with the consensus score was .54 and the partial standardized coefficient without modeling measurement error was .51; compared to the dissattenuated estimate (.69) this result highlights how measurement error can distort coefficients. The AEIM confidence scores correlated less strongly with personality and intelligence but what these scores measure is rather unclear.

For assessing the incremental validity of the AEIM scores, we first added intelligence and the personality traits to predict the outcomes; these controls were jointly significant, which stresses the

importance of keeping theoretically relevant variables in the model. We then entered the AEIM scores, finding that none of their coefficients or the joint *F*-tests for the *r*-square change of the AEIM scores were significant. This result is not surprising given that the AEIM's concurrent (i.e., correlation with the outcomes) and discriminant (i.e., overlap with intelligence and empathy) validities were weak. As a robustness check, we also included an interaction between the AEIM consensus and confidence scores, while accounting for measurement error (see Busemeyer & Jones, 1983) and boostrapping standard errors. This interaction might be considered a methodologically sound alternative to looking at high and low combinations of the two AEIM measures. The coefficient of the interaction was insignificant and substantive results were unchanged.

Our results are reminiscent of those by Schulte, Ree, and Carretta (2004), who showed that the MSCEIT ability EI test had a reliability-corrected multiple correlation of .81 with the big five, intelligence, and gender. We found a slightly lower multiple correlation (without gender because these data were not available to us). These results and others (Amelang & Steinmayr, 2006) cast doubt on the validity of EI ability tests. Specifically, our results show that the AEIM should not be used in applied settings for assessment or to predict performance.

Conclusion

Our analyses show that the use of stepwise regression, the forming of subgroups with extreme scores analyses, and ignoring heteroscedasticity and measurement error can result in flawed validity tests. These statistical procedures compromise the scientific method at the expense of maximizing the likelihood of reporting validity when there is none.

References

Amelang, M., & Steinmayr, R. (2006). Is there a validity increment for tests of emotional intelligence in explaining the variance of performance criteria? *Intelligence, 34*(5), 459-468.

Ando, M., & Hodoshima, J. (2007). A note on bootstrapped White's test for heteroskedasticity in regression models. *Economics Letters, 97*(1), 46-51.

Antonakis, J., Ashkanasy, N. M., & Dasborough, M. T. (2009). Does leadership need emotional intelligence? *The Leadership Quarterly, 20*(2), 247-261.

Antonakis, J., & Dietz, J. (2010). Emotional intelligence: On definitions, neuroscience, and marshmallows. *Industrial and Organizational Psychology, 3*(2), 165-170.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica, 47*, 1287-1294.

Busemeyer, J. R., & Jones, L. E. (1983). Analysis of Multiplicative Combination Rules When the Causal Variables Are Measured with Error. *Psychological Bulletin, 93*(3), 549-562.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. New York: Cambridge University Press.

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society Series B-Methodological, 45*(3), 311-354.

Davies, M., Stankov, L., & Roberts, R. D. (1998). Emotional Intelligence: In Search of an Elusive Construct. *Journal of Personality and Social Psychology, 75*(4), 989-1015.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: Wiley.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica, 47*(1), 153-161.

Leigh, J. P. (1988). Assessing the importance of an independent variable in multiple-regression--Is stepwise unwise? *Journal of Clinical Epidemiology, 41*(7), 669-677.

Locke, E. A. (2005). Why emotional intelligence is an invalid concept. *Journal of Organizational Behavior, 26*(4), 425-431.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113*(1), 181-190.

Meehl, P. E. (1959). Some ruminations on the validation of clinical procedures. *Canadian Journal of Psychology, 13*, 102-128.

Ree, M. J., & Carretta, T. R. (2006). The role of measurement error in familiar statistics. *Organizational Research Methods, 9*(1), 99-112.

Schmidt, F. L., & Hunter, J. E. (1999). Theory Testing and Measurement Error. *Intelligence, 27*(3), 183-198.

Schulte, M. J., Ree, M. J., & Carretta, T. R. (2004). Emotional Intelligence: Not much more than g and personality. *Personality and Individual Differences, 37*(5), 1059-1068.

Sechrest, L. (1963). Incremental validity - A recommendation. *Educational and Psychological Measurement, 23*(1), 153-158.

Tavris, C., & Aronson, E. (2007). *Mistakes were made (but not by me): Why we justify foolish beliefs, bad decisions, and hurtful acts*. New York: Harcourt.

Thompson, B. (1995). Stepwise Regression and Stepwise Discriminant-Analysis Need Not Apply Here - A Guidelines Editorial. *Educational and Psychological Measurement, 55*(4), 525-534.

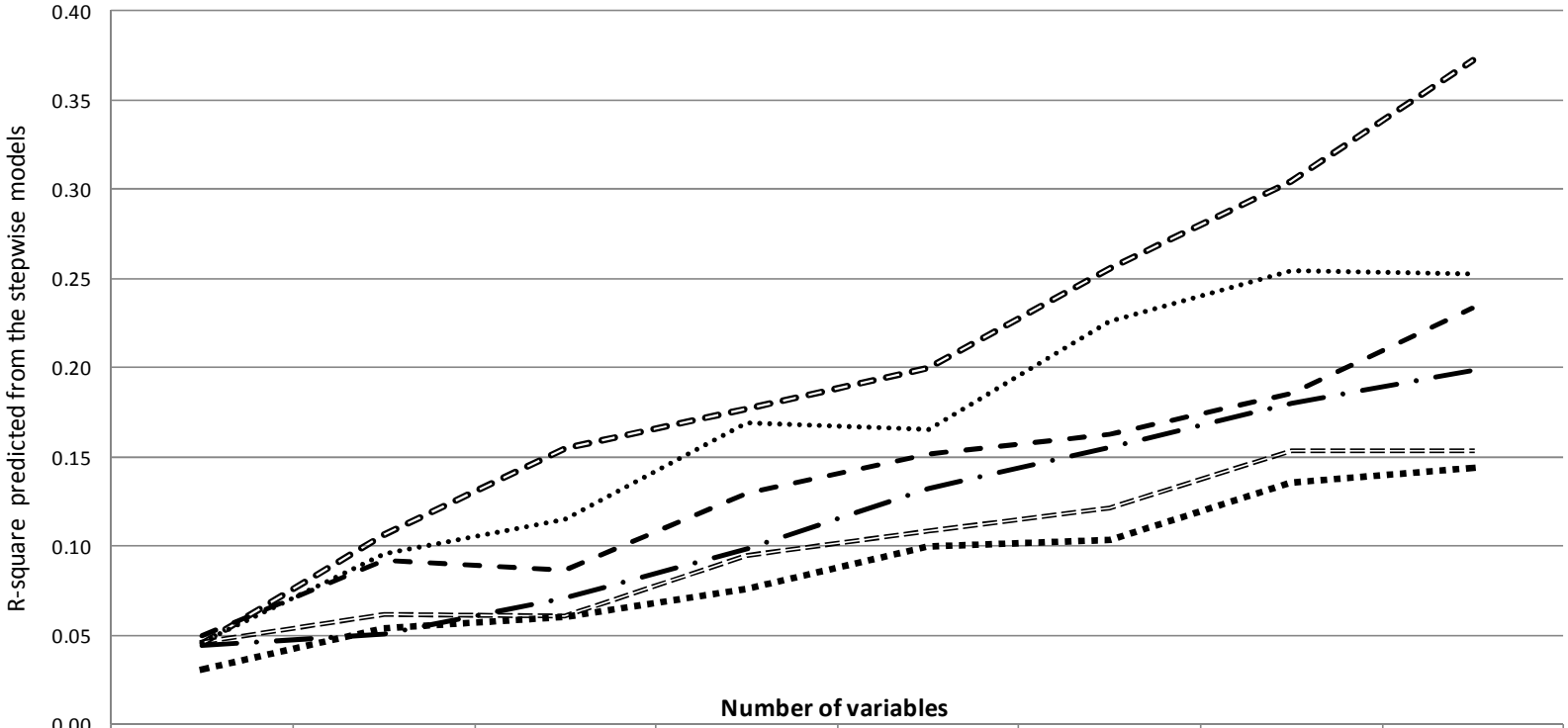
Tobin, J. (1958). Estimation of relationships for limited dependent variables. Econometrica, 26, 24-36.

Warwick, J., Nettelbeck, T., & Ward, L. (2010). AEIM: A new measure and method of scoring abilities-based emotional intelligence. *Personality and Individual Differences, 48*(1), 66-71.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*, 817-830.

Zeidner, M., Roberts, R. D., & Matthews, G. (2008). The science of emotional intelligence: Current consensus and controversies. *European Psychologist, 13*(1), 64-78.

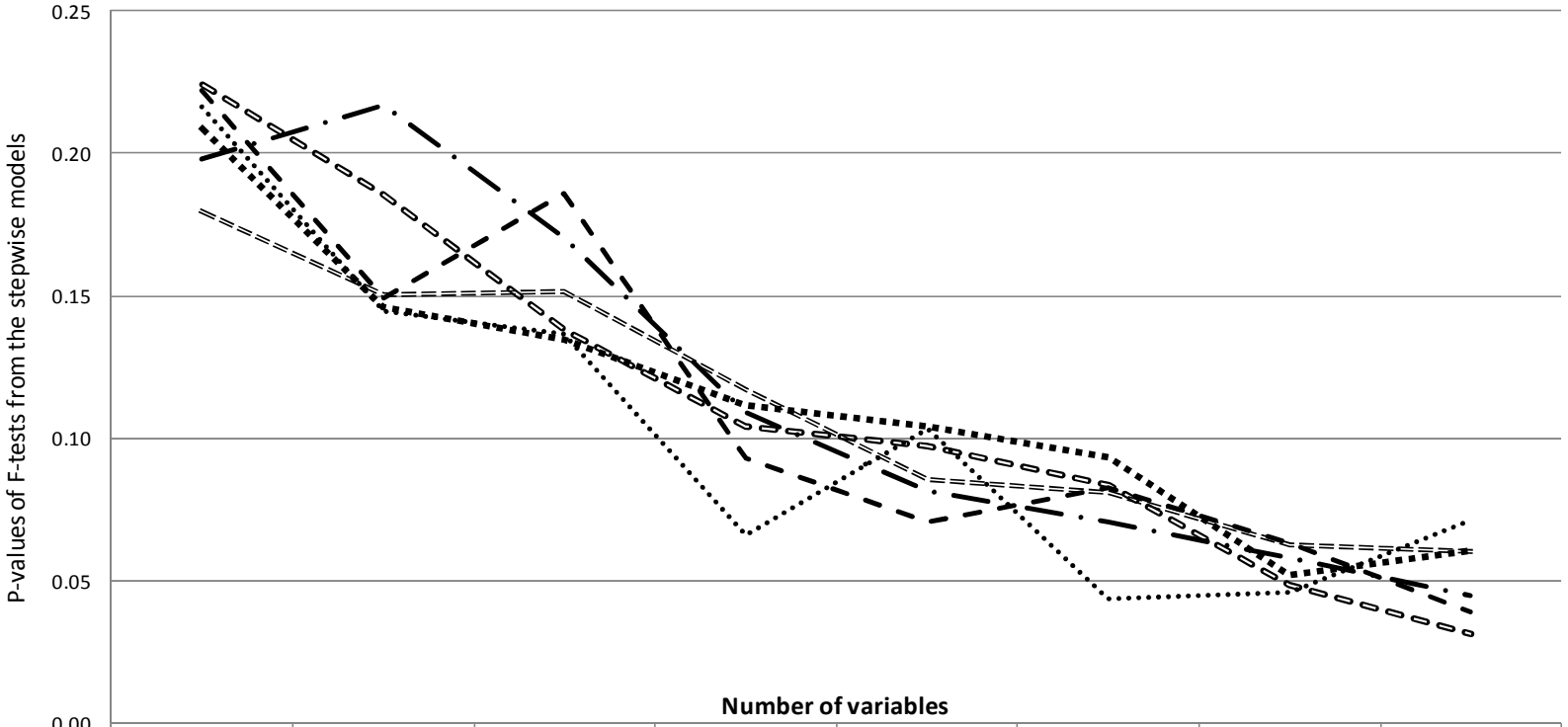
Figure 1: Stepwise Regression Simulations for R-Squares



Sample size	2	3	4	5	6	7	8	9
20	0.05	0.11	0.16	0.18	0.20	0.26	0.30	0.37
25	0.05	0.10	0.11	0.17	0.17	0.23	0.25	0.25
30	0.05	0.09	0.09	0.13	0.15	0.16	0.19	0.23
35	0.04	0.05	0.07	0.10	0.13	0.16	0.18	0.20
40	0.05	0.06	0.06	0.09	0.11	0.12	0.15	0.15
45	0.03	0.05	0.06	0.08	0.10	0.10	0.14	0.14

Note: Using a Monte Carlo simulation ($k = 20$ replications), we varied the sample size and number of predictors, which, with *y*, were pure "noise" variables (table entries are the *r*-square values depicted in the figure).

Figure 2: Stepwise Regression Simulations for p-values of F-tests

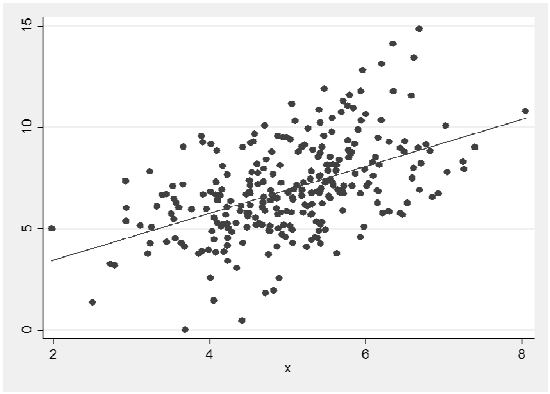


Sample size	2	3	4	5	6	7	8	9
20	0.22	0.19	0.14	0.10	0.10	0.08	0.05	0.03
25	0.22	0.14	0.14	0.07	0.10	0.04	0.05	0.07
30	0.22	0.15	0.19	0.09	0.07	0.08	0.06	0.04
35	0.20	0.22	0.17	0.11	0.08	0.07	0.06	0.04
40	0.18	0.15	0.15	0.12	0.09	0.08	0.06	0.06
45	0.21	0.15	0.13	0.11	0.10	0.09	0.05	0.06

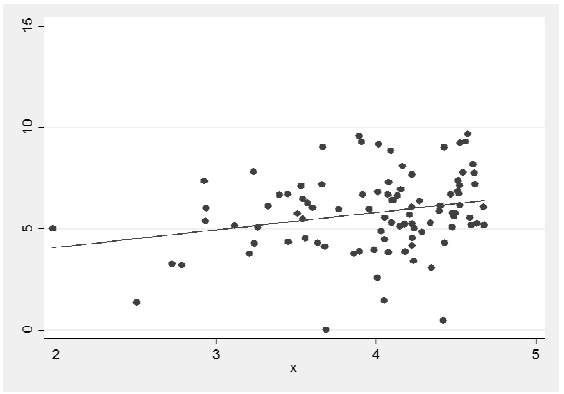
Note: *p*-values of the *F*-tests for the stepwise regression in Figure 1 (table entries are *p*-values depicted in the figure). *F*-tests for the regression equations become significant as a function of the number of variables included in the model

.

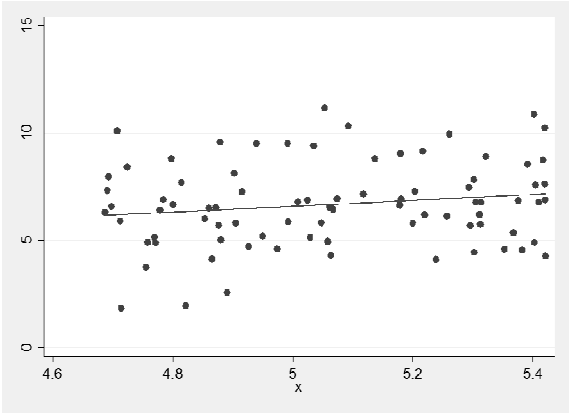
Figure 3: Attenuation of Relations with Subgroup



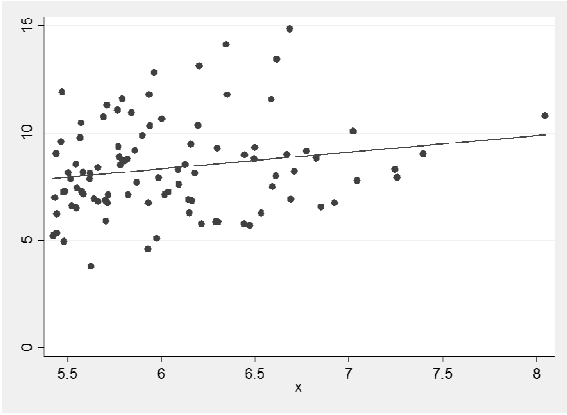
Full sample (1)



Low sample (2)



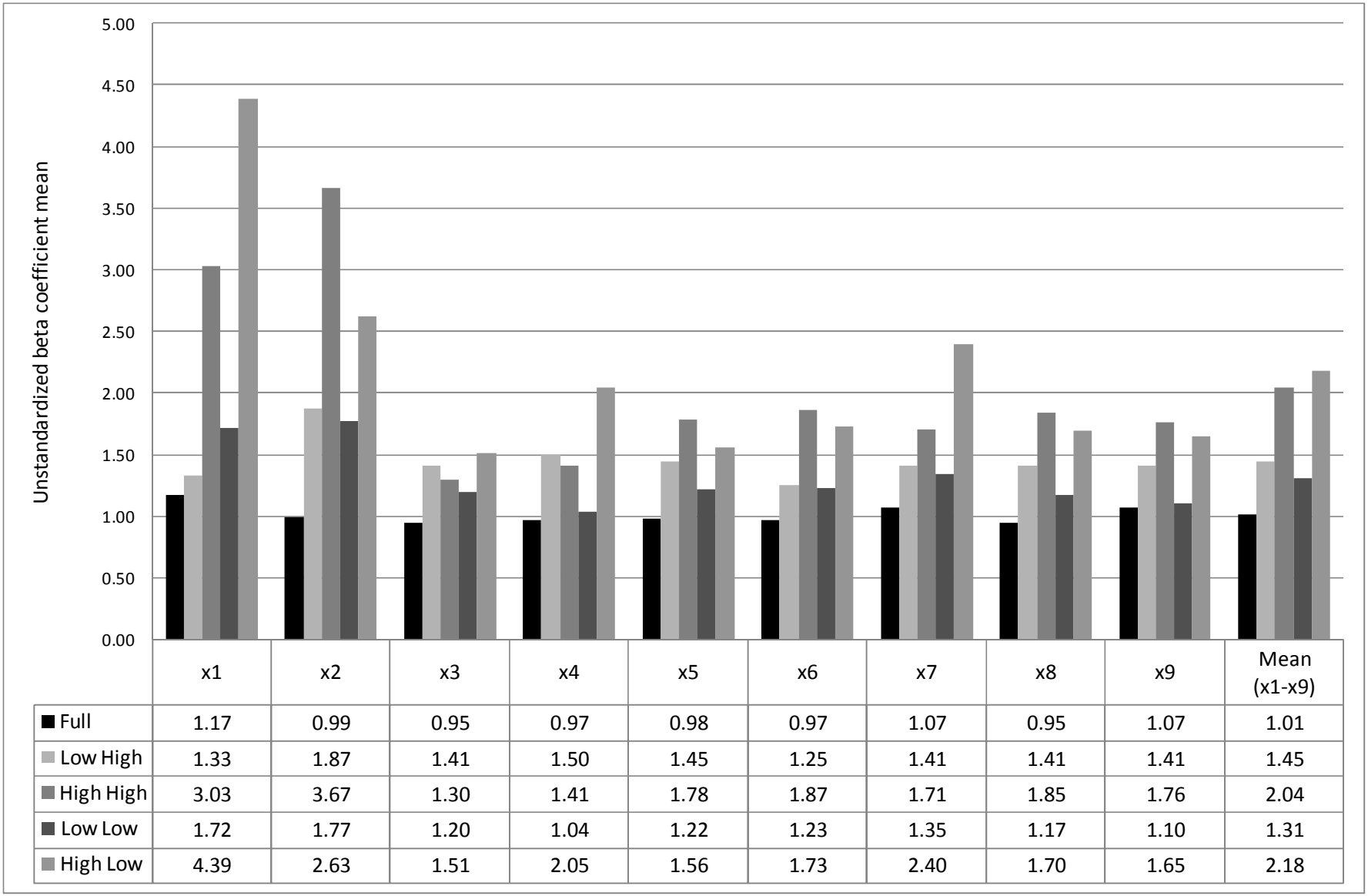
Middle sample (3)



High sample (4)

Note: This figure shows how subgrouping a sample (n = 272) attenuates relations not only in the middle range of the sample but also in the extremes (graphs from one simulation). We then used a Monte Carlo simulation (k = 20 replications) to show that in the population, the relation between x (horizontal axes) and y (vertical axes) is r = .44; however, in the "low" sample it was r = .23, in the "middle" sample it was r = .14, and in the "high" sample it was r = .24. These values were all different from the population value, t-statistics (df = 19) being -9.27, -11.04, and -10.05 respectively (p's < .001, two-tailed). Estimating a regression model across sample 2 and 4 only would actually *accentuate* the relationship: r = .50, which is also different from the population value, t =4.30, p < .01 (two-tailed).

Figure 4: Effect of Subgrouping and Stepwise Regression



	x1	x2	x3	x4	x5	x6	x7	x8	x9	Mean (x1-x9)
Full	1.17	0.99	0.95	0.97	0.98	0.97	1.07	0.95	1.07	1.01
Low High	1.33	1.87	1.41	1.50	1.45	1.25	1.41	1.41	1.41	1.45
High High	3.03	3.67	1.30	1.41	1.78	1.87	1.71	1.85	1.76	2.04
Low Low	1.72	1.77	1.20	1.04	1.22	1.23	1.35	1.17	1.10	1.31
High Low	4.39	2.63	1.51	2.05	1.56	1.73	2.40	1.70	1.65	2.18

Note: This Monte Carlo experiment (20 replications) compares parameter estimates using the full sample and all variables (with regression analysis) with that of sub-grouping analysis (i.e., combinations of the top and bottom 35%) using stepwise regression. We split into four groups using cut-offs on $x1$ and $x2$ (to retain the low and high combinations of $x1$ and $x2$). Low-High, High-High, Low-Low, and High-Low refer to the group splitting combinations. With stepwise, the mean beta coefficients for the four groups were severely overestimated (between 31%-118%); however, the mean of the full sample using regression analysis was, at 1.01 (at the correct value). The number of regressors retained in the four respective groups was 5, 4, 7, and 3, and the r-squares were .71, .39, .67, and .47; however, with the full sample (including all nine regressors, which were always significant), the *r*-square was .50. Note, using normal regression and all controls within subgroups still produced inaccurate findings. n = 272.

Table 1: Correlation Matrix and Descriptive Statistics (Warwick et al. data)

	Mean	S.D.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1 EI consensus	.52	.07	*.70*													
2 EI confidence	82.86	9.79	.36	*.96*												
3 IQ	28.25	3.73	.54	.33	*.82*											
4 IQ Fluid	28.14	5.03	.52	.32	.82	*.90*										
5 IQ Verbal	28.36	4.42	.33	.19	.76	.24	*.73*									
6 Extraversion	27.18	5.91	.07	.08	-.07	-.03	-.10	*.80*								
7 Neuroticism	21.63	6.40	-.18	-.15	-.17	-.15	-.11	-.40	*.77*							
8 Openness	29.96	4.32	.09	.15	.14	.09	.14	.19	.03	*.60*						
9 Conscient.	29.71	4.97	.07	.15	.12	.12	.06	.03	.01	.14	*.82*					
10 Agreeable.	28.59	4.02	.14	.06	.04	.00	.06	-.07	.14	.20	.37	*.60*				
11 Empathy	12.38	1.84	.28	.14	.13	.18	.01	.02	.18	.32	.20	.36	.84			
12 GPA	67.93	9.75	.13	.07	.25	.11	.30	-.22	.05	.10	.10	-.03	.10	-		
13 Stress	28.39	7.18	-.16	-.07	-.12	-.10	-.08	-.18	.54	-.05	.03	-.06	.09	-.05	-	
14 Lonely	41.20	9.65	-.23	-.17	-.14	-.19	-.03	-.49	.55	.04	-.12	-.06	-.08	-.02	.50	-
15 Well-being	11.52	2.37	.11	.09	.08	.12	.00	.30	-.44	.00	.22	.10	.00	.10	-.34	-.62

Note: For indicative purposes (and with the caveat that the *t* statistics are biased because of the heteroscedasticity problem), $p < .10$ for *r's* >|.10|; $p < .05$ for *r's* > |.12|;

/cont.

$p < .01$ for *r's* > |.16|; $p < .001$for *r's* >|.22|; N=272; alpha reliabilities on the diagonal; because Warwick et al. did not report the reliability for the IQ scale (based on all fluid and verbal IQ items), we approximated it to be the mean of the fluid and verbal IQ scores; not knowing the reliabilities of the last four variables (GPA to Well-being) is actually irrelevant when these variables are modeled as dependent variables given that the measurement error is pooled in the error term of the equation and does not affect the consistency of estimates of the independent variables.

Table 2: Errors-in-variables Regression Models (AEIM data)

Model	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)	(5a)	(5b)	(6a)	(6b)
Variables	EI Cons.	EI Cons.	EI Conf.	EI Conf.	GPA	GPA	Stress	Stress	Lonely	Lonely	Well-being	Well-being
EI Cons.					.08	.06	-.03	-.03	.08	.08	-.08	-.07
					(.61)	(.47)	(-.24)	(-.32)	(.78)	(.77)	(-.71)	(-.63)
EI Conf.					-.05	-.07	.09	.08	-.05	-.05	-.03	-.03
					(-.38)	(-.58)	(.65)	(.63)	(-.46)	(-.47)	(-.24)	(-.20)
IQ Fluid	.46***		.25**		-.11		-.04		-.15*		.11	
	(3.57)		(2.44)		(-.83)		(-.34)		(-1.78)		(1.10)	
IQ Verb.	.32***		.13		.33***		.12		-.05		-.06	
	(2.90)		(1.21)		(2.60)		(.90)		(-.44)		(-.57)	
IQ Full		.69***		.34***		.16		.05		-.18		.05
		(4.42)		(2.89)		(1.02)		(.39)		(-1.36)		(.39)
Extraver.	.15	.18	.01	.03	-.38**	-.42**	.23	.22	-.48***	-.49***	.15	.17
	(1.26)	(1.34)	(.07)	(.24)	(-2.13)	(-2.42)	(1.49)	(1.58)	(-3.28)	(-3.48)	(1.03)	(1.10)
Neurot.	-.12	-.09	-.15	-.13	-.07	-.08	.89***	.89***	.53***	.52***	-.55***	-.54***
	(-1.09)	(-.73)	(-1.10)	(-.95)	(-.42)	(-.51)	(6.35)	(6.49)	(3.94)	(4.00)	(-4.90)	(-4.52)

/cont.

Openn.	-.20	-.24	.12	.10	.22	.29	-.21	-.18	.43**	.45**	-.14	-.17
	(-1.37)	(-1.50)	(.87)	(.68)	(1.10)	(1.46)	(-1.14)	(-1.14)	(2.37)	(2.55)	(-.86)	(-1.03)
Consc.	-.15	-.15	.11	.11	.26	.22	.20	.18	.02	.02	.16	.17
	(-1.41)	(-1.45)	(1.01)	(1.01)	(1.63)	(1.51)	(1.45)	(1.37)	(.19)	(.15)	(1.17)	(1.32)
Agreeab.	.25	.25	-.04	-.04	-.48	-.40	-.41*	-.38*	-.32	-.31	.27	.25
	(1.37)	(1.37)	(-.20)	(-.24)	(-1.52)	(-1.44)	(-1.86)	(-1.77)	(-1.62)	(-1.51)	(1.38)	(1.27)
Empathy	.25**	.26***	.09	.11	.20	.12	.11	.09	-.22*	-.23**	.03	.06
	(2.37)	(2.81)	(.82)	(.98)	(1.13)	(.82)	(.92)	(.77)	(-1.93)	(-2.13)	(.20)	(.51)
Constant	.05	.01	43.40***	39.54***	51.17***	6.18***	5.72	8.02	68.83***	71.29***	1.71***	9.81***
	(.53)	(.08)	(3.21)	(2.85)	(3.06)	(3.85)	(.54)	(.81)	(5.27)	(5.78)	(3.29)	(2.83)
R-square	.45	.48	.17	.18	.26	.20	.48	.47	.59	.59	.35	.34
Mult. R	.67	.69	.41	.42	.51	.45	.69	.69	.77	.77	.59	.58

Note: *** p<0.01, ** p<0.05, * p<0.10; Each model is estimated twice: Once (a) with IQ as one general factor and once (b) as two factors (fluid & verbal IQ); parameter estimates are standardized; numbers in parentheses are *z* statistics from normal bootstrapped standard errors (findings regarding the AEIM were unchanged when using percentile or bias-corrected bootstraps); N=272.