Alexander Weiss1 and Paul T. Costa, Jr.2,3
(e-mail: alex.weiss@ed.ac.uk)

1 Department of Psychology, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh EH8 9JZ, United Kingdom
2 Behavioral Medicine Research Center, Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, NC 27710
3 Department of Mental Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205

DOI: 10.1093/aje/kwt300; Advance Access publication: February 5, 2014

THE AUTHORS REPLY

We thank Drs. Weiss and Costa for their letter (1), in which they suggest that the limited statistical power of individual cohort studies is not ameliorated by a meta-analysis that pools the results of several cohort studies together. We find this argument surprising given that higher statistical power is one of the main advantages of meta-analysis (2). For example, in our meta-analysis of 3,947 deaths in 76,150 participants from 7 cohort studies (3), the power to detect a hazard ratio of 0.88 for an association between conscientiousness and mortality increased from 42% when the meta-analysis included only 2 cohort studies—the British Household Panel Survey and the Household, Income, and Labour Dynamics in Australia Survey—to 98% when 5 additional cohort studies were added to the random-effects meta-analysis. In separate proportional hazards models, the 7 individual cohort studies had, in order of decreasing numbers of deaths, 90% power to correctly reject the null hypothesis, representing hazard ratios of 0.90, 0.89, 0.88, 0.86, 0.84, 0.83, and 0.82 per standard-deviation difference in personality score (3). Thus, our meta-analysis had sufficient power to detect even small associations.

Weiss and Costa also note that the 7 cohort studies did not have identical personality scales, which might have led to “content and construct deficiency” of the traits (1, p. 791). However, the instruments used in our meta-analysis were not very heterogeneous. Two of the cohort studies used the full 44-item Big Five Inventory; 2 used the short 15-item Big Five Inventory, which is derived from the 44-item inventory; 2 used the 26-item Midlife Development in the United States (MIDUS) personality inventory; and 1 used the 35-item Character Trait Descriptive Adjectives inventory.

All of these personality inventories were developed using appropriate psychometric analyses. There is also evidence of considerable convergent validity between the different instruments used to assess the 5-factor traits (4, 5). For example, John et al. (4) reported correlations ($r$) of 0.80 between corresponding traits measured by the 44-item Big Five Inventory and the 100-item Character Trait Descriptive Adjectives inventory, 0.77 between the Big Five Inventory and the 60-item NEO Five-Factor Inventory, and 0.75 between the Character Trait Descriptive Adjectives inventory and the NEO Five-Factor Inventory.

We agree that predictive validity may be weaker for brief personality scales than for longer and more detailed scales (6), but this difference is unlikely to “seriously compromise” the predictive validity of current measures (5, 6) as Weiss and Costa suggested (1, p. 791). Our meta-analysis did not include cohort studies with very brief 5- or 10-item inventories, which may be particularly vulnerable to reduced predictive validity (6). Using several of the same cohort studies as those employed here, we have previously shown that low conscientiousness was also the only 5-factor personality trait associated with the development of obesity (7) and adult-onset diabetes (8). The risk of coronary heart disease was additionally predicted by high neuroticism and the risk of stroke by high extraversion (9). These results support the importance of conscientiousness but also suggest that the other 5-factor traits may be relevant for specific health outcomes.

Our current meta-analysis does not exclude the possibility that extraversion, neuroticism, agreeableness, and openness to experience might be associated with a very modest increased risk of all-cause mortality, that other personality constructs not measured in this meta-analysis are associated with mortality, or that associations may exist in some subgroups that were not included in our supplementary analyses. Nonetheless, our null findings from a large pooled data set cast doubt about the importance of extraversion, neuroticism, agreeableness, and openness to experience as major psychological risk factors for premature mortality.

ACKNOWLEDGMENTS

The authors were financially supported by the Academy of Finland (grants 124322, 124271, and 132944); the Bupa Foundation, United Kingdom; the Medical Research Council (grant K013351); the National Institutes of Health (grants R01HL036310 and R01AG034454); the New Occupational Safety and Health ERA
In their letter, Weiss and Costa (1) rightfully decry the paucity of statistical power arising from low mortality rates in the Jokela et al. (2) data sets. I agree that the null overall results Jokela et al. report for 4 of the Big 5 personality traits are inconclusive for this reason. Instead, the heterogeneity in risk estimates for other dimensions of personality hints at either imprecision arising from low mortality rates, unknown effect modifiers in different samples, or some of both. Perhaps the best interpretation, with which I suspect Jokela et al. would agree, is that conscientiousness has a larger/more consistent association with mortality than do the other Big 5. The remaining 4 axes of personality variation are not irrelevant to survival but are probably important in their own ways among particular persons within particular environments at particular times of life.

Weiss and Costa’s argument about measurement is a rather more complex issue, to which I am sympathetic (3). The emergence of integrative data analysis in psychology has focused on the psychometric harmonization of measurement instruments across samples (4). This issue also arises in genetic epidemiology when the same phenotype is measured differently across disparate studies, which then must be pooled to achieve adequate power. However, the harmonization of psychometric measures across samples, usually done with item response theory, requires a different approach.

In the absence of such a sample, Jokela et al. were left with the eternal “apples and oranges” dilemma of meta-analysis and had only conceptual grounds to guide them. I think they fared rather well in this regard, because the “Big Five Inventory” in some of their data sets was actually developed to reconcile Big 5 questionnaire measures with trait adjective measures—one version of which is used in Jokela et al.’s other data sets (5). Thus, the two major measurement vehicles across studies were historically developed to converge. Nevertheless, there is an important distinction between “semantic equivalence” (i.e., identical names of scales) and “construct equivalence” (scales measuring the same unobserved variable), and it is difficult to evaluate the latter directly, as Weiss and Costa note (1).

The larger issue that Weiss and Costa raise is that the Big 5 are very complex, multifaceted constructs, and they are measured with exceeding simplification in the Jokela et al. data sets. In essence, the Big 5 were developed as higher-order dimensions in a hierarchically nested taxonomy of personality variation. It is therefore dif- ficult to specify the particular persons within particular environments at particular times of life.

In their letter, Weiss and Costa (1) rightfully decry the paucity of statistical power arising from low mortality rates in the Jokela et al. (2) data sets. I agree that the null overall results Jokela et al. report for 4 of the Big 5 personality traits are inconclusive for this reason. Instead, the heterogeneity in risk estimates for other dimensions of personality hints at either imprecision arising from low mortality rates, unknown effect modifiers in different samples, or some of both. Perhaps the best interpretation, with which I suspect Jokela et al. would agree, is that conscientiousness has a larger/more consistent association with mortality than do the other Big 5. The remaining 4 axes of personality variation are not irrelevant to survival but are probably important in their own ways among particular persons within particular environments at particular times of life.

Weiss and Costa’s argument about measurement is a rather more complex issue, to which I am sympathetic (3). The emergence of integrative data analysis in psychology has focused on the psychometric harmonization of measurement instruments across samples (4). This issue also arises in genetic epidemiology when the same phenotype is measured differently across disparate studies, which then must be pooled to achieve adequate power. However, the harmonization of psychometric measures across samples, usually done with item response theory, requires a “linkage sample” which contains data on all scales to be rendered commensurate. In the absence of such a sample, Jokela et al. were left with the eternal “apples and oranges” dilemma of meta-analysis and had only conceptual grounds to guide them. I think they fared rather well in this regard, because the “Big Five Inventory” in some of their data sets was actually developed to reconcile Big 5 questionnaire measures with trait adjective measures—one version of which is used in Jokela et al.’s other data sets (5). Thus, the two major measurement vehicles across studies were historically developed to converge. Nevertheless, there is an important distinction between “semantic equivalence” (i.e., identical names of scales) and “construct equivalence” (scales measuring the same unobserved variable), and it is difficult to evaluate the latter directly, as Weiss and Costa note (1).

The larger issue that Weiss and Costa raise is that the Big 5 are very complex, multifaceted constructs, and they are measured with exceeding simplification in the Jokela et al. data sets. In essence, the Big 5 were developed as higher-order dimensions in a hierarchically nested taxonomy of personality variation. Just as a Linnaean genus contains widely variegated


DR. CHAPMAN REPLIES

In their letter, Weiss and Costa (1) rightfully decry the paucity of statistical power arising from low mortality rates in the Jokela et al. (2) data sets. I agree that the null overall results Jokela et al. report for 4 of the Big 5 personality traits are inconclusive for this reason. Instead, the heterogeneity in risk estimates for other dimensions of personality hints at either imprecision arising from low mortality rates, unknown effect modifiers in different samples, or some of both. Perhaps the best interpretation, with which I suspect Jokela et al. would agree, is that conscientiousness has a larger/more consistent association with mortality than do the other Big 5. The remaining 4 axes of personality variation are not irrelevant to survival but are probably important in their own ways among particular persons within particular environments at particular times of life.

Weiss and Costa’s argument about measurement is a rather more complex issue, to which I am sympathetic (3). The emergence of integrative data analysis in psychology has focused on the psychometric harmonization of measurement instruments across samples (4). This issue also arises in genetic epidemiology when the same phenotype is measured differently across disparate studies, which then must be pooled to achieve adequate power. However, the harmonization of psychometric measures across samples, usually done with item response theory, requires a “linkage sample” which contains data on all scales to be rendered commensurate. In the absence of such a sample, Jokela et al. were left with the eternal “apples and oranges” dilemma of meta-analysis and had only conceptual grounds to guide them. I think they fared rather well in this regard, because the “Big Five Inventory” in some of their data sets was actually developed to reconcile Big 5 questionnaire measures with trait adjective measures—one version of which is used in Jokela et al.’s other data sets (5). Thus, the two major measurement vehicles across studies were historically developed to converge. Nevertheless, there is an important distinction between “semantic equivalence” (i.e., identical names of scales) and “construct equivalence” (scales measuring the same unobserved variable), and it is difficult to evaluate the latter directly, as Weiss and Costa note (1).

The larger issue that Weiss and Costa raise is that the Big 5 are very complex, multifaceted constructs, and they are measured with exceeding simplification in the Jokela et al. data sets. In essence, the Big 5 were developed as higher-order dimensions in a hierarchically nested taxonomy of personality variation. Just as a Linnaean genus contains widely variegated


DR. CHAPMAN REPLIES

In their letter, Weiss and Costa (1) rightfully decry the paucity of statistical power arising from low mortality rates in the Jokela et al. (2) data sets. I agree that the null overall results Jokela et al. report for 4 of the Big 5 personality traits are inconclusive for this reason. Instead, the heterogeneity in risk estimates for other dimensions of personality hints at either imprecision arising from low mortality rates, unknown effect modifiers in different samples, or some of both. Perhaps the best interpretation, with which I suspect Jokela et al. would agree, is that conscientiousness has a larger/more consistent association with mortality than do the other Big 5. The remaining 4 axes of personality variation are not irrelevant to survival but are probably important in their own ways among particular persons within particular environments at particular times of life.

Weiss and Costa’s argument about measurement is a rather more complex issue, to which I am sympathetic (3). The emergence of integrative data analysis in psychology has focused on the psychometric harmonization of measurement instruments across samples (4). This issue also arises in genetic epidemiology when the same phenotype is measured differently across disparate studies, which then must be pooled to achieve adequate power. However, the harmonization of psychometric measures across samples, usually done with item response theory, requires a “linkage sample” which contains data on all scales to be rendered commensurate. In the absence of such a sample, Jokela et al. were left with the eternal “apples and oranges” dilemma of meta-analysis and had only conceptual grounds to guide them. I think they fared rather well in this regard, because the “Big Five Inventory” in some of their data sets was actually developed to reconcile Big 5 questionnaire measures with trait adjective measures—one version of which is used in Jokela et al.’s other data sets (5). Thus, the two major measurement vehicles across studies were historically developed to converge. Nevertheless, there is an important distinction between “semantic equivalence” (i.e., identical names of scales) and “construct equivalence” (scales measuring the same unobserved variable), and it is difficult to evaluate the latter directly, as Weiss and Costa note (1).

The larger issue that Weiss and Costa raise is that the Big 5 are very complex, multifaceted constructs, and they are measured with exceeding simplification in the Jokela et al. data sets. In essence, the Big 5 were developed as higher-order dimensions in a hierarchically nested taxonomy of personality variation. Just as a Linnaean genus contains widely variegated