HHS Public Access

Author manuscript

Psychosom Med. Author manuscript; available in PMC 2020 June 01.

Published in final edited form as:

Psychosom Med. 2019 June ; 81(5): 408–414. doi:10.1097/PSY.0000000000000701.

Between the Error Bars: How Modern Theory, Design, and Methodology Enrich the Personality-Health Tradition

Suzanne C. Segerstrom, Ph.D., M.P.H.

Department of Psychology, University of Kentucky, 125 Kastle Hall, Lexington, KY 40506-0044, Phone 859-257-4549, FAX 859-323-1979, segerstrom@uky.edu

Abstract

The study of relationships between personality traits and health has a long history in psychosomatic research. However, personality science has evolved from an understanding of personality as fixed traits to one that acknowledges that personality is dynamic. Dynamic approaches to conceptualizing and measuring personality and individual differences can enrich personality-health research. In this Presidential Address (American Psychosomatic Society, 2018), I consider how different formulations of personality – stable traits, stable signals in a noisy or variable measure, within-person changes, and intraindividual variability – can be implemented to better understand how personality is related to health and particularly to immune function. These approaches recognize and, in some cases, capitalize on the fact that personality factors can display variability as well as stability over time. They also require repeated measurement and therefore greater methodological sophistication that considers reliability and generalizability, Simpson's paradox, and the difference between variability and flexibility. Dynamic qualities of personality and individual differences potentially influence health, and designs and methodology that incorporate them can illuminate the important processes that occur inside the error bars.

Keywords

psychoneuroimmunology; cortisol; optimism; repetitive thought; intraindividual variability; methodology

1. Personality and health

The relationship between personality and health has a long and distinguished history in the study of psychosomatic relationships. This body of work includes documentation of the health effects of "Type" A personality and its active ingredient, hostility; dispositional optimism; neuroticism and negative affectivity; and others. One assumption has been that these constructs are trait-like: they show substantial temporal stability and are therefore effectively captured by a single assessment. There is evidence to support this assumption. Test-retest reliabilities indicate that adults who are high on a trait at one time are also likely to be high on that trait years later (e.g., hostility over 3 years, $r = 0.84$ [1]; dispositional optimism over 10 years, $r = 0.35$ [2]; neuroticism over 6 years, $r = 0.83$ [3]). Rank-order stability for temperament and personality in childhood and adolescence is lower but still

nontrivial (4). Second, correlations between one-time assessments and prospective health outcomes suggest that such assessments have predictive validity, even when personality was assessed in childhood (5).

Additional approaches to conceptualizing and measuring personality and individual differences can enrich this tradition. These approaches recognize and, in some cases, capitalize on the fact that personality and individual differences are capable of variability as well as stability. First, one can consider personality as the consistent thread through a measure that appears unstable: the “signal” in the “noise”. Affect and cortisol, for example, are notoriously variable, but average levels could reflect an underlying trait. Second, one can take the “noise” seriously. Changes (i.e., instability) in individual differences over time could correlate with changes in health. Finally, some people fluctuate to a greater degree than others, and these people can be said to have more intraindividual or within-person variability. Like other traits, intraindividual variability has its own distribution in the population and can predict important health outcomes. In this review, I consider how these different formulations of personality – traits, “signals”, unstable individual differences, and intraindividual variability – can be implemented to better understand how personality is related to health and particularly to immune function. I illustrate them by drawing on my own work and its evolution from focusing on traits to explicitly incorporating variability into the study of personality and health.

2. Traits and immune function

My interest in individual differences arose from influential studies of stress and immune function published in the 1980s (e.g., 6,7). These studies demonstrated changes in immune parameters as a function of psychological stress, but my attention was drawn to the error bars. Clearly, not everyone experienced the same degree of immune change as a function of stress. What could explain these differences?

My dissertation was designed to test whether optimism – the expectation of a positive future – could protect against stress-related immune change (8). First-year law students completed questionnaires measuring their general expectancies for the future (dispositional optimism) and their expectancies specifically about law school (situational optimism) and had blood drawn before they started school and again two months later. Dispositional optimism before starting school was generally unrelated to change in immune parameters over the two months. (Effects of dispositional optimism on immune parameters turned out to be strongly determined by circumstances [9,10].) Higher situational optimism, however, was associated with more T cells and higher natural killer (NK) cell cytotoxicity during the first semester of law school after adjusting for baseline levels.

I have continued to pursue effects of stable individual differences on physiology and health. A more recent study investigated immunological correlates of resources, the evaluation of one’s possession of financial status, social connections, psychological strengths, and personal skills (11). Among both young women (mean age = 27 years) and older women (mean age = 82 years), more perceived resources correlated with a lower percentage of senescent NK cells ($CD3^{-}CD56^{dim}CD57^{+}$). The pool of NK cells contains more and more

of these senescent cells with age, but the difference between having low and high resources was nearly as large as the difference between young and old adulthood. Among the older women, one standard deviation difference in resources was equivalent to 5 years of age. Individual differences inside the error bars can be substantive and potentially important for health.

3. Signal and noise: Methodological considerations

Perceived resources, which predicted NK cell subsets in the study described above, are a stable individual difference. This is easy to see in Figure 1, which shows the distribution of resource scores when measured annually and repeatedly in older adults. The individual with values in green consistently reported high resources, the individual with values in red consistently reported lower resources, and the individual in blue reported the lowest resources, except at one time point when he or she reported higher resources.

It is more difficult to discern whether there are stable individual differences in the data shown in Figure 2. This figure illustrates the distribution of diurnal cortisol slopes (computed from 4 observations between waking and 9 pm), measured repeatedly on first-year law students. Presumably, some students had consistently steeper slopes than others, but the amount of variation makes it difficult to discern whether the three people whose daily slopes are shown in green, red, and blue differ from each other.

Salivary cortisol can be a useful measure. First, it is the endpoint of the hypothalamic-pituitary-adrenal axis, an important system for stress response and energy regulation. Second, it is easy to collect (not requiring venipuncture). Third, it represents the unbound, bioactive fraction of cortisol, and therefore is physiologically significant. Finally, a large literature demonstrates that salivary cortisol is responsive to psychological states. It appears to be an ideal measure for studying psychosomatic phenomena, and probably hundreds of studies have employed it (12–15).

However, when not implemented thoughtfully, salivary cortisol is not a useful measure. Cortisol, particularly a derived measure such as diurnal slope, varies substantially within individuals from day to day. On any given day, only 10–20% of the variance in salivary cortisol slope is due to stable individual differences, so there can be little signal in relation to a great deal of noise (16–18). A generalizability and decision study (18, 19) examined how many sampling days would yield reliable measures of diurnal slope and area under the curve. Figure 3 illustrates the results from the law students shown in Figure 2. For area under the curve, minimal reliability (0.60) would be expected around 4 days of sampling and good reliability (0.80), around 10 days. For diurnal slope, minimal reliability would be expected around 10 days of sampling, and good reliability would require over 20 days. Similar results were obtained for older adults.

The consequences of failing to adequately address reliability and generalizability can be profound. Most scientists know that low reliability can result in Type II error, failing to detect a statistically significant relationship where a true relationship exists, as random error obscures the relationship. Low reliability can also result in Type I error, detecting a

statistically significant relationship where no true relationship exists; because low reliability introduces random error, unreliable measurement can erroneously produce a relationship just as randomly generated data can also erroneously produce a relationship (20).

However, Type I and II errors are based on *p* values, which are relatively uninformative. When decisions are made about what findings to pursue and with what sample size and power or when the clinical significance of a finding is evaluated, effect size, not *p* value, is informative. Type S (sign) and Type M (magnitude) errors are cases in which the point estimate of an effect size is inaccurate (21). Sign errors occur when an estimated effect has the opposite sign from the underlying true effect, and magnitude errors occur when an estimated effect is meaningfully larger or smaller than the underlying true effect. Design and methodology are critical to the rates of these errors. “When researchers use small samples and noisy measurement to study small effects – as they often do in psychology as well as other disciplines – a significant result is often surprisingly likely to be in the wrong direction [S error] and to greatly overestimate an effect [M error]” (21, p. 641).

The combination of a small sample, noisy measurement, and small effect size is not uncommon, and the empirical literature on psychosocial influences on cortisol seems a particularly vivid example. In a meta-analysis of psychological health and the cortisol waking response, we identified 186 independent effects (15). This literature as a whole has characteristics associated with S and M errors. The meta-analytic effect size was $r = .09$ (small), and the median sample size was 74 (ergo, at least half of the studies had small sample sizes). About 30% of the variance in the cortisol waking response arises from stable individual differences. Therefore, depending on the exact measure (e.g., slope versus area under the curve), reliable measurement requires 2–6 days of sampling (22). In the meta-analysis, the median number of days of cortisol sampling was 2 (mode = 1), so measurements were almost certainly noisy in many studies.

Data from 24 macaques illustrate the possible consequences of noisy measurement (23). Handedness (which has been associated with individual differences in behavior and immune function in monkeys; 23) was assessed with a reach test, and blood was drawn for the assessment of NK cell cytotoxicity (NKCC) every month for a year. Generalizability for one measurement was poor (0.29) but for the average of all measurements was good (0.84). When the most reliable average – of all 13 measurements – was used, there was a statistically significant relationship between right-handedness and higher NKCC ($b = 12.7$, $SE = 4.5$, $p < .01$). That relationship and its 95% confidence interval are at the far right of Figure 4.

At the far left of Figure 4 is the range of 1000 beta weights obtained when each monkey had one NKCC value of the 13 chosen at random, and the regression was re-run. The average beta weight over the 1000 regression models was very close to the “true” relationship, but the range was very large, and there were both S errors (negative beta weights, below the solid line) and M errors (values outside the 95% CI of the best measurement model at right, outside the dotted lines). Clearly, noisy measurement can result in spurious results. When two NKCC values were chosen at random for each monkey and averaged (yielding generalizability of 0.45), there were fewer sign errors but still many magnitude errors. The

average of 4 values (0.62) generally resulted in few errors, and the average of 10 values (0.80) resulted in estimates that were closely centered on the best estimate.

A frequent question when people see data such as these is, can a large sample size overcome noisy measurement? That is, can noisy measurement and a small effect size still give a reliable result with a large enough sample size? This question is particularly pertinent for parameters requiring sampling that would be unrealistic for most investigators and participants (e.g., more than 20 days' sampling for a good estimation of trait salivary cortisol slope.)

Figure 5 shows the results from a simulation study to answer that question. In this example case, the true relationship between *x* and *y* (both normally distributed) was a standardized beta weight of 0.10 (i.e., a small effect size), and the measurement reliability of *y* was 0.30 (i.e., noisy, as might be obtained from the average of three days' cortisol slopes in law students or one measurement of NKCC in a macaque), 0.60 (i.e., adequate), or 0.80 (i.e., good). Ten thousand datasets were generated for each sample size of 50, 100, 250, 750, and 1500. When measurement was noisy (reliability = 0.30), small error rates did not occur until sample size exceeded 1000. The answer to the question, then, is that large sample sizes can overcome noisy measurement, but the sample must be very large if reliability is low and the effect size is small. When measurement was more reliable, errors decreased more rapidly with increasing sample size. Note that increasing the effect size decreased the number of sign errors, but not the number of magnitude errors. (See the Supplemental Digital Content for the SAS code used to generate these datasets. The interested reader can use the supplied code to test any combination of effect size, sample size, and measurement reliability of Y and to sample from different distributions of X and Y.)

The trait "signal" in an inherently noisy measure can be located but may pose challenges, including large sample sizes and demanding methodology, if misestimation is to be avoided. Diurnal cortisol provides a good example case, but other biomarkers may pose similar challenges related to their "physiometrics" (24).

4. Unstable individual differences: Making use of the noise

Rather than trying to overcome the noise in an inherently variable measure, it is possible to turn those unstable individual differences to one's advantage by testing the effects of within-person change. For example, optimism about law school varied over the first six months, and these changes in optimism were relevant for cellular immunity. When first-year students had situational optimism measured along with delayed-type hypersensitivity (DTH) tests (a measure of cellular immunity) at 5 time points, within-person increases in optimism correlated with increases in DTH (26).

Although the within-person analysis did provide a conceptual replication of the findings in my dissertation on optimism and immunity (see 2. Above), the assumption that differences between people must be the same as changes within people is erroneous. In fact, individual differences and changes can even have opposite effects. Take, for example, the relationship between exercise and heart attack risk. When comparing different people, individuals who

exercise more have lower heart attack risk than those who exercise less. However, for any given person, that person is more likely to have a heart attack when exercising than when at rest. Between-person differences in exercise and within-person changes in exercise have opposite effects. This phenomenon is known as Simpson's paradox or the ecological fallacy (27). Between-person relationships can only be generalized to within-person relationships under strict requirements of homogeneity across people and time (ergodicity; summarized in 28). These requirements are unlikely to be met in many cases. One commonly sees generalizations from between-person differences (e.g., people who are mindful are healthier) to within-person changes (e.g., people who start to practice mindfulness will become healthier). Because of Simpson's paradox, between-person effects provide poor evidence for within-person effects.

In the domain of self-rated health (SRH), for example, between-person differences in affect and within-person changes in affect interacted differently with age to predict SRH (29). Between people, average negative affect was unrelated to SRH among the young-old (e.g., 60-year-olds), but less negative affect was associated with better SRH among the old-old (e.g., 90-year-olds). This interaction was different for within-person change: decreases in negative affect were associated with better SRH among the young-old, but changes in affect were unrelated to SRH among the old-old. Treating within-person changes as substantive individual differences rather than noise allows one to disentangle these effects (30).

Another example emerges from work on the dimensional model of repetitive thought (RT). In this model, discrete forms of RT (e.g., worry, rumination, planning) are characterized along three orthogonal dimensions: Valence (thoughts about positive content vs. thoughts about negative content); Purpose (searching, uncertain, questioning thoughts vs. solving, certain, asserting thoughts); and Total (high endorsement of many kinds of RT vs. low endorsement). This structure replicated in both undergraduate students and older adults using various batteries of RT measures (31–33). The dimensional model has the advantage of being capable of isolating the active ingredients of RT. For example, worry has high Total, negative Valence, and solving Purpose. Assessment of worry alone does not allow one to ask which component is responsible for increased risk of cardiovascular disease (34). In the dimensional model, Valence and Total had different correlates in psychological well-being (Valence correlated with well-being, whereas Total correlated with psychological growth; 35), neuropsychological function (Valence correlated with executive function, whereas Total correlated with verbal intelligence; 32), mindfulness (Valence correlated with nonreactivity, whereas Total correlated with nonjudging; 33), and interleukin-6 (Valence was uncorrelated, whereas Total was negatively correlated; 36). These different correlates of RT dimensions suggest that complex, heterogeneous RT types, such as worry, do not provide causal unity, and "When it occurs that a previously recognized psychological construct is subdivided into more elemental components that have different etiologies, or different external correlates, or that require different interventions, it no longer makes sense to treat the original entity as a coherent, homogeneous construct and to represent it by a single score" (37, p. 273; see also 38).

However, the dimensional model was difficult to implement because it required an entire battery of RT measures. To overcome this barrier, we developed and validated an 8-item

circumplex method (39). With this less burdensome method, it became possible to examine both between-person differences and within-person changes in RT dimensions. As part of the validation work, undergraduates responded to the 8 RT items and the Center for Epidemiological Studies-Depression scale every day for up to 14 days. Averages across the 14 days replicated findings with trait RT: higher average negative valence was the strongest correlate of more depressive symptoms, with higher average total RT also associated, but less strongly. Changes in RT dimensions from day to day, however, did not mirror this pattern. First, valence was less influential: negative valence, searching purpose, and high total had similar correlations with depressive symptoms. (The three dimensions are orthogonal to one another, and interactions among the dimensions did not correlate with depressive symptoms.) Therefore, the within-person correlates of depressive symptoms were not the same as the between-person correlates, an illustration of Simpsons' paradox.

More interesting, both the valence and total dimensions had random slopes. That is, the relationship between changes in RT dimensions and changes in depressive symptoms differed across people. For valence, some people had strong correlations between RT and symptoms, but other people had almost no correlation at all. For total, some people experienced more depressive symptoms on days when they had more RT, but other people experienced *fewer* depressive symptoms.

In general, theoretical and empirical work on RT focuses on exposure, the amount of a particular kind of RT in which a person engages. For example, the widely used Response Styles Questionnaire asks, "When you feel sad, down, or depressed, how *often* do you …" (emphasis added). However, by taking the "noise" of RT seriously, we found that individual differences in reactivity – that is, affective responsiveness to RT – is also important. Two people endorsing the same kind or amount of RT may have very different affective experiences accompanying that thought. Indeed, RT could be considered an internally generated event, and like external events, both reactivity and exposure may be important for the psychological and physical consequences (40, 41).

In summary, changes in optimism, affectivity, or repetitive thought can be converted from random noise that one tries to overcome in a stable-trait model to substantive predictive variance in an unstable-trait model. Doing so has the advantage of providing evidence of the effects of within-person change, evidence that is potentially different – and more clinically pertinent – than that of the effects of between-person differences.

5. Stable individual differences in within-person change

The final method of using noise to advantage involves individual differences in the magnitude of within-person change. Re-examining Figure 2, although it is difficult to discern mean differences among the three people whose values are highlighted, it is evident that the person whose values are shown in green was more variable over time than the person in red, and the person in blue was the most consistent over time. Figure 6 shows the distribution of individual standard deviations (iSD) in diurnal cortisol slope in the entire sample of first-year law students (i.e., each student's standard deviation of his or her observations over up to 15 days).

Affect, reaction time, life satisfaction, and other psychological phenomena as well as cortisol show these individual differences in intraindividual variability (IIV). (For more examples, see the December 2009 special issue of *Psychology and Aging* or the August 2017 special issue of *Journal of Research in Personality*.) How does one capture that variability, and is it important? As was true of the mean of repeated measurements, reliable measurement of the iSD of repeated measurements requires a sufficient number of observations, and that number is in fact higher than the corresponding number to reliably capture the mean (for mathematical derivation, see ref. 42; for other methodological considerations in using the iSD, see refs. 43–45).

More important, IIV has substantive implications for health. In the Midlife in the United States sample, larger iSDs in negative affect were associated both cross-sectionally and prospectively with worse physical health (a composite of self-rated health, chronic conditions, limitations in activities of daily living, and number of prescription medications) (46). In general, higher IIV in psychological and cognitive measures has predicted higher morbidity and earlier mortality (46–48) as well as correlating with poorer psychological health (46, 50, 51). More variability may reflect more volatility (e.g., higher affective IIV) and less cognitive control (e.g., higher reaction time IIV).

However, more variability might not always be bad. In a collaboration with Julia Boehm, Ashley Winning, and Laura Kubzansky (52), we found that IIV in life satisfaction moderated the relationship between mean life satisfaction and risk of premature mortality. For people who were volatile (i.e., they had larger iSDs in life satisfaction over 9 annual assessments), those with generally low life satisfaction had the highest risk of premature mortality, but those with generally high life satisfaction had the lowest risk. By contrast, people who were stable (i.e., they had smaller iSDs in life satisfaction) had similar risk regardless of their general level of life satisfaction. In this study, IIV might have indicated not undesirable volatility but rather sensitivity. Like “orchid” children (53), people with high IIV in life satisfaction might be demonstrating reactivity to their environments; and like “dandelion” children, people with low IIV might be showing that they are relatively insensitive.

Finally, variability is not the same as flexibility. Jaime Hardy and I defined variability as “the range or frequency of a response, uncharacterized by situational change” and flexibility as “the ability to vary one’s responses in a contextually dependent manner in order to appropriately meet situational demands” (37, p. 13). Unlike affective variability, affective flexibility (as in Zautra’s dynamic model of affect; 54) was associated with better psychological and physical health. Similarly, variability of cortisol slopes in the law students was associated with lower neuroticism. We concluded that “A high slope iSD does not imply disruption of the diurnal rhythm of cortisol across the day . … One way of interpreting the present finding is that people higher in neuroticism had less flexibility in cortisol regulation, which could be maladaptive in a system that is designed to be sensitive to environmental demands.” (46, p. 123) Intraindividual variability that is responsive to situational change may suggest adaptive flexibility; in contrast, inherent variability may suggest a lack of system integrity in some cases (e.g., reaction time; 49) but certainly not all (e.g., heart rate; 55).

6. Conclusion

The relationship between personality and health has a long and distinguished history, dating back at least to the 1940s in *Psychosomatic Medicine* (56–58). Since then, personality science has evolved from an understanding of personality as fixed traits to one that acknowledges that personality is dynamic. Stable traits such as neuroticism and extraversion nonetheless have substantial IIV (59), variable constructs such as affect can be influenced by stable tendencies, and variability itself is a form of personality. All of these dynamic qualities potentially influence health, and designs and methodology that incorporate them into research on personality and health can illuminate the important processes that occur inside the error bars.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Conflicts of Interest and Source of Funding:

The author has no conflicts to declare. Preparation of this manuscript was supported by the National Institute on Aging (K02–033629).

Glossary

NK	natural killer
NKCC	natural killer cell cytotoxicity
DTH	delayed-type hypersensitivity
SRH	self-rated health
RT	repetitive thought
iSD	individual standard deviation
IIV	intraindividual variability

References

(1). Shekelle RB, Gale M, Ostfeld AM, Paul O. Hostility, risk of coronary heart disease, and mortality. Psychosom Med 1983;45:109–14. [PubMed: 6867229]

(2). Segerstrom SC. Optimism and resources: Effects on each other and on health over 10 years. J Res Pers 2007;41:772–86.

(3). Costa PT, McCrae RR. Personality in adulthood: a six-year longitudinal study of self-reports and spouse ratings on the NEO Personality Inventory. J Pers Soc Psychol 1988;54:853–63. [PubMed: 3379583]

(4). Roberts BW, DelVecchio WF. The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies. Psychol Bull 2000;126:3–25. [PubMed: 10668348]

(5). Friedman HS, Tucker JS, Tomlinson-Keasey C, Schwartz JE, Wingard DL, Criqui MH. Does childhood personality predict longevity? J Pers Soc Psychol 1993;65:176–85. [PubMed: 8355139]

(6). Irwin M, Daniels M, Smith TL, Bloom E, Weiner H. Impaired natural killer cell activity during bereavement. Brain Beh Immun 1987;1:98–104.

(7). Kiecolt-Glaser JK, Glaser R, Strain EC, Stout JC, Tarr KL, Holliday JE, Speicher CE. Modulation of cellular immunity in medical students. J Behav Med 1986;9:5–21. [PubMed: 2939253]

(8). Segerstrom SC, Taylor SE, Kemeny ME, Fahey JL. Optimism is associated with mood, coping, and immune change in response to stress. J Pers Soc Psychol 1998;74:1646–55. [PubMed: 9654763]

(9). Segerstrom SC. Optimism and immunity: Do positive thoughts always lead to positive effects? Brain Beh Immun 2005;19:195–200.

(10). Segerstrom SC. How does optimism suppress immunity? Evaluation of three affective pathways. Health Psychol 2006;25:653–7. [PubMed: 17014284]

(11). Segerstrom SC, Al-Attar A, Lutz CT. Psychosocial resources, aging, and natural killer cell terminal maturity. Psychol Aging 2012;27:892–902. [PubMed: 22708535]

(12). Denson TF, Spanovic M, Miller N. Cognitive appraisals and emotions predict cortisol and immune responses: a meta-analysis of acute laboratory social stressors and emotion inductions. Psychol Bull 2009;135:823–53. [PubMed: 19883137]

(13). Chida Y, Steptoe A. Cortisol awakening response and psychosocial factors: a systematic review and meta-analysis. Biol Psychol 2009;80:265–78. [PubMed: 19022335]

(14). Boggero IA, Hostinar CE, Haak EA, Murphy ML, Segerstrom SC. Psychosocial functioning and the cortisol awakening response: Meta-analysis, P-curve analysis, and evaluation of the evidential value in existing studies. Biol Psychol 2017;129:207–30. [PubMed: 28870447]

(15). Knorr U, Vinberg M, Kessing LV, Wetterslev J. Salivary cortisol in depressed patients versus control persons: a systematic review and meta-analysis. Psychoneuroendocrinol 2010;35:1275–86.

(16). Doane LD, Chen FR, Sladek MR, Van Lenten SA, Granger DA. Latent trait cortisol (LTC) levels: Reliability, validity, and stability. Psychoneuroendocrinol 2015;55:21–35.

(17). Ross KM, Murphy ML, Adam EK, Chen E, Miller GE. How stable are diurnal cortisol activity indices in healthy individuals? Evidence from three multi-wave studies. Psychoneuroendocrinol 2014;39:184–93.

(18). Segerstrom SC, Boggero IA, Smith GT, Sephton SE. Variability and reliability of diurnal cortisol in younger and older adults: implications for design decisions. Psychoneuroendocrinol 2014;49:299–309.

(19). Webb NM, Shavelson RJ, Haertel EH. Four reliability coefficients and generalizability theory. Handbook Stat 2006;26:81–124.

(20). Duffy S Random numbers demonstrate the frequency of Type I errors: three spreadsheets for class instruction. J Stat Educ 2010;18(2).

(21). Gelman A, Carlin J. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. Persp Psychol Sci 2014;9:641–51.

(22). Hellhammer J, Fries E, Schweisthal OW, Schlotz W, Stone AA, Hagemann D. Several daily measurements are necessary to reliably assess the cortisol rise after awakening: state-and trait components. Psychoneuroendocrinol 2007;32:80–6.

(23). Segerstrom SC, Lubach GR, Coe CL. Identifying immune traits and biobehavioral correlates: generalizability and reliability of immune responses in rhesus macaques. Brain Beh Immun 2006;20:349–58.

(24). Segerstrom SC, Smith G. Methods, variance, and error in psychoneuroimmunology research: the good, the bad, and the ugly In: Segerstrom SC, editor. The Oxford handbook of psychoneuroimmunology. New York: Oxford; 2012, p. 421–432.

(25). Schönbrodt FD, Perugini M. At what sample size do correlations stabilize? J Res Pers 2013;47:609–12.

(26). Segerstrom SC, Sephton SE. Optimistic expectancies and cell-mediated immunity: The role of positive affect. Psychol Sci 2010;21:448–55. [PubMed: 20424083]

(27). Kievit R, Frankenhuis WE, Waldorp L, Borsboom D. Simpson’s paradox in psychological science: a practical guide. Frontiers Psychol 2013;4:513.

(28). Molenaar PC, Campbell CG. The new person-specific paradigm in psychology. Curr Dir Psychol Sci 2009;18:112–7.

(29). Segerstrom SC. Affect and self-rated health: A dynamic approach with older adults. Health Psychol 2014;33:720–8. [PubMed: 23914813]

(30). Ryu E, West SG, Sousa KH. Distinguishing between-person and within-person relationships in longitudinal health research: Arthritis and quality of life. Ann Behav Med 2012;43:330–42. [PubMed: 22270265]

(31). Segerstrom SC, Stanton AL, Alden LE, Shortridge BE. A multidimensional structure for repetitive thought: What's on your mind, and how, and how much? J Pers Soc Psychol 2003;85:909–21. [PubMed: 14599253]

(32). Segerstrom SC, Roach AR, Evans DR, Schipper LJ, Darville AK. The structure and health correlates of trait repetitive thought in older adults. Psychol Aging 2010;25:505–15. [PubMed: 20677888]

(33). Evans DR, Segerstrom SC. Why do mindful people worry less? Cognit Ther Res 2011;35:505–10.

(34). Kubzansky LD, Kawachi I, Spiro A III, Weiss ST, Vokonas PS, Sparrow D. Is worrying bad for your heart? A prospective study of worry and coronary heart disease in the Normative Aging Study. Circulation 1997;95:818–24. [PubMed: 9054737]

(35). Segerstrom SC, Eisenlohr-Moul TA, Evans DR, Ram N. Repetitive thought dimensions, psychological well-being, and perceived growth in older adults: a multilevel, prospective study. Anx Stress Coping 2015;28:287–302.

(36). Segerstrom SC, Reed RG, Scott AB. Intelligence and interleukin-6 in older adults: the role of repetitive thought. Psychosom Med 2017;79:757–62. [PubMed: 28445209]

(37). Smith GT, McCarthy DM, Zapolski TC. On the value of homogeneous constructs for construct validation, theory testing, and the description of psychopathology. Psychol Assess 2009;21(3): 272–84. [PubMed: 19719340]

(38). Mõttus R Towards more rigorous personality trait–outcome research. Eur J Pers 2016;30:292–303.

(39). Segerstrom SC, Hardy JK, Evans DR, Boggero IA, Alden LE, Stanton AL. Briefly assessing repetitive thought dimensions: valence, purpose, and total. Assessment 2016;23:614–23. [PubMed: 26019299]

(40). Bolger N, Schilling EA. Personality and the problems of everyday life: The role of neuroticism in exposure and reactivity to daily stressors. J Pers 1991;59:355–86. [PubMed: 1960637]

(41). Suls J, Martin R. The daily life of the garden-variety neurotic: Reactivity, stressor exposure, mood spillover, and maladaptive coping. J Pers 2005;73:1485–510. [PubMed: 16274443]

(42). Wang L, Grimm KJ. Investigating reliabilities of intraindividual variability indicators. Multivar Behav Res 2012;47:771–802.

(43). Baird BM, Le K, Lucas RE. On the nature of intraindividual personality variability: Reliability, validity, and associations with well-being. J Pers Soc Psychol 2006;90:512–27. [PubMed: 16594835]

(44). Segerstrom SC, Sephton SE, Westgate PM. Intraindividual variability in cortisol: Approaches, illustrations, and recommendations. Psychoneuroendocrinol 2017;78:114–24.

(45). Wang LP, Hamaker E, Bergeman CS. Investigating inter-individual differences in short-term intra-individual variability. Psychol Methods 2012;17:567–81. [PubMed: 22924600]

(46). Hardy J, Segerstrom SC. Intra-individual variability and psychological flexibility: Affect and health in a National US sample. J Res Pers 2017;69:13–21.

(47). Eizenman DR, Nesselroade JR, Featherman DL, Rowe JW. Intraindividual variability in perceived control in a older sample: The MacArthur successful aging studies. Psychol Aging 1997;12:489–502. [PubMed: 9308096]

(48). Batterham PJ, Bunce D, Mackinnon AJ, Christensen H. Intra-individual reaction time variability and all-cause mortality over 17 years: a community-based cohort study. Age Ageing 2014;43:84–90. [PubMed: 23934546]

(49). Tales A, Leonards U, Bompas A, Snowden RJ, Philips M, Porter G, Haworth J, Wilcock G, Bayer A. Intra-individual reaction time variability in amnestic mild cognitive impairment: a precursor to dementia? J Alzheimers Dis 2012;32:457–66. [PubMed: 22785393]

(50). Eid M, Diener E. Intraindividual variability in affect: Reliability, validity, and personality correlates. J Pers Soc Psychol 1999;76:662–76.

(51). Kuppens P, Van Mechelen I, Nezlek JB, Dossche D, Timmermans T. Individual differences in core affect variability and their relationship to personality and psychological adjustment. Emotion 2007;7:262–74. [PubMed: 17516805]

(52). Boehm JK, Winning A, Segerstrom S, Kubzansky LD. Variability modifies life satisfaction's association with mortality risk in older adults. Psychol Sci 2015;26:1063–70. [PubMed: 26048888]

(53). Ellis BJ, Boyce WT. Biological sensitivity to context. Curr Dir Psychol Sci 2008;17:183–7.

(54). Zautra A, Smith B, Affleck G, Tennen H. Examinations of chronic pain and affect relationships: Applications of a dynamic model of affect. J Consult Clin Psychol 2001;69:786–95. [PubMed: 11680555]

(55). Malik M, Camm AJ. Heart rate variability. Clin Cardiol 1990;13:570–6. [PubMed: 2204508]

(56). Mittelmann B, Weider A, Vonachen HA, Kronenberg M, Weider NO, Brodman KE, Wolff HG. Detection and management of personality and psychosomatic disorders among industrial personnel. Psychosom Med 1945;7:359–67. [PubMed: 21005003]

(57). Brozek J, Guetzkow H, Keys A, Cattell RB, Harrower MR, Hathaway SR. A study of personality of normal young men maintained on restricted intakes of vitamins of the B complex. Psychosom Med 1946;8:98–109. [PubMed: 21019919]

(58). Brodman K, Mittelmann B, Wechsler D, Weider A, Wolff HG, Meixner MD. The relation of personality disturbances to duration of convalescence from acute respiratory infections. Psychosom Med 1947;9(1):37–44. [PubMed: 20284395]

(59). Fleeson W Toward a structure-and process-integrated view of personality: Traits as density distributions of states. J Pers Soc Psychol 2001;80:1011–27. [PubMed: 11414368]

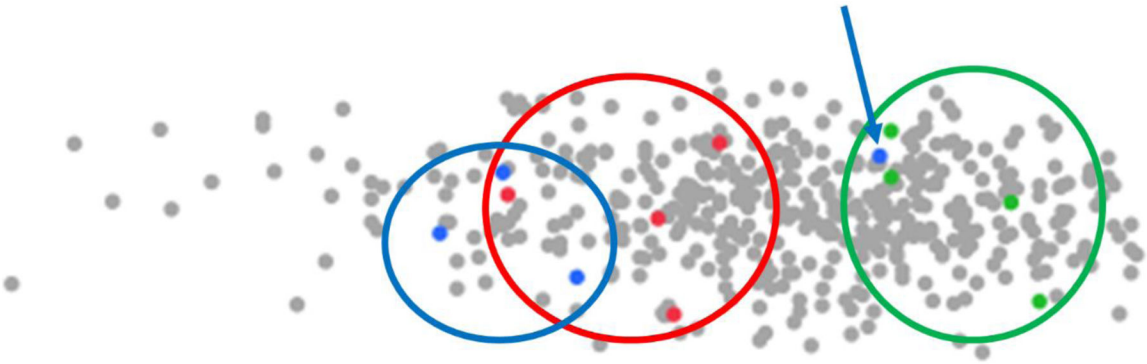

Figure 1.
The distribution of resource scores (lowest at left, highest at right; observations are staggered vertically to improve visibility) among 150 older adults. Three individuals' scores, measured annually, are highlighted and circled in green, red, and blue.

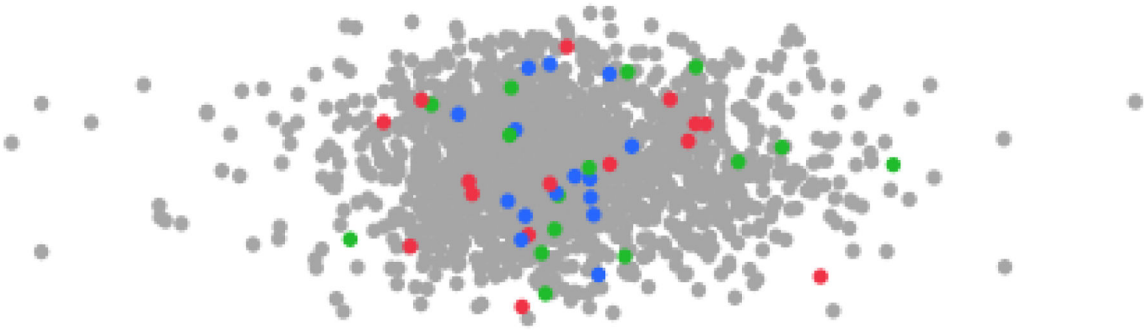

Figure 2.

The distribution of daily diurnal cortisol slopes (lowest at left, highest at right; observations are staggered vertically to improve visibility) among 125 law students. Three individuals' slopes, measured up to 15 times over 6 months, are highlighted in green, red, and blue.

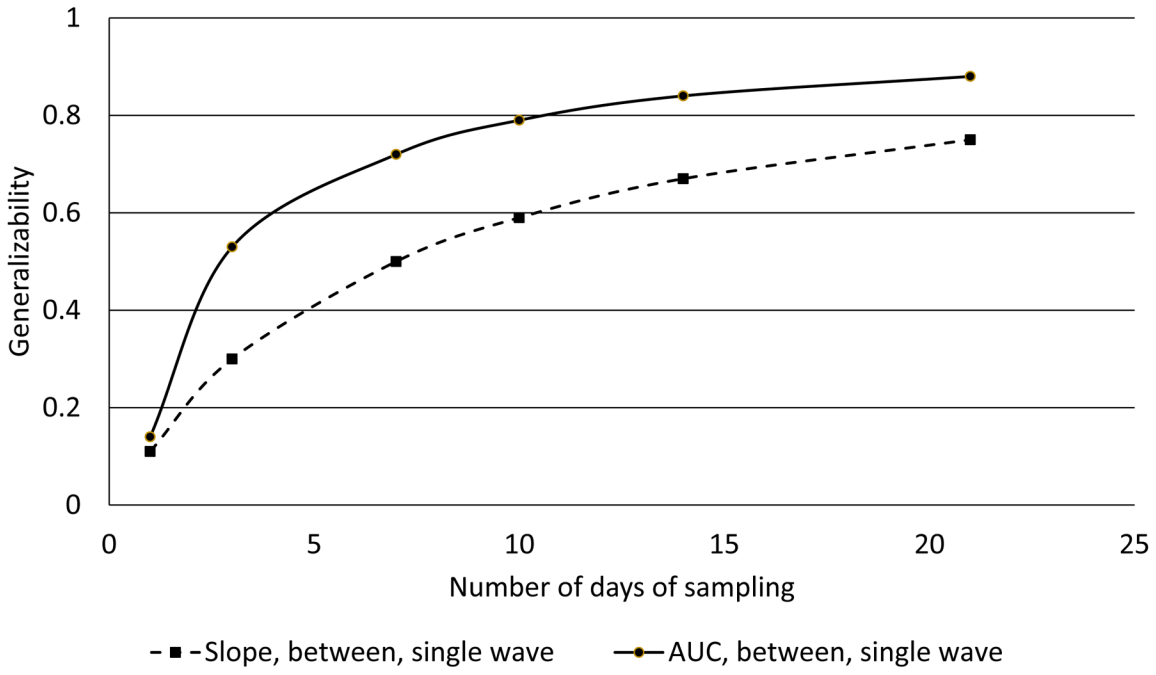

Figure 3.

Results of a decision study for diurnal cortisol slope and area under the curve. Number of sampling days is on the X axis; generalizability (the expected percentage of variance that is "true" variance) for stable between-person differences is on the Y axis. AUC = area under the curve.

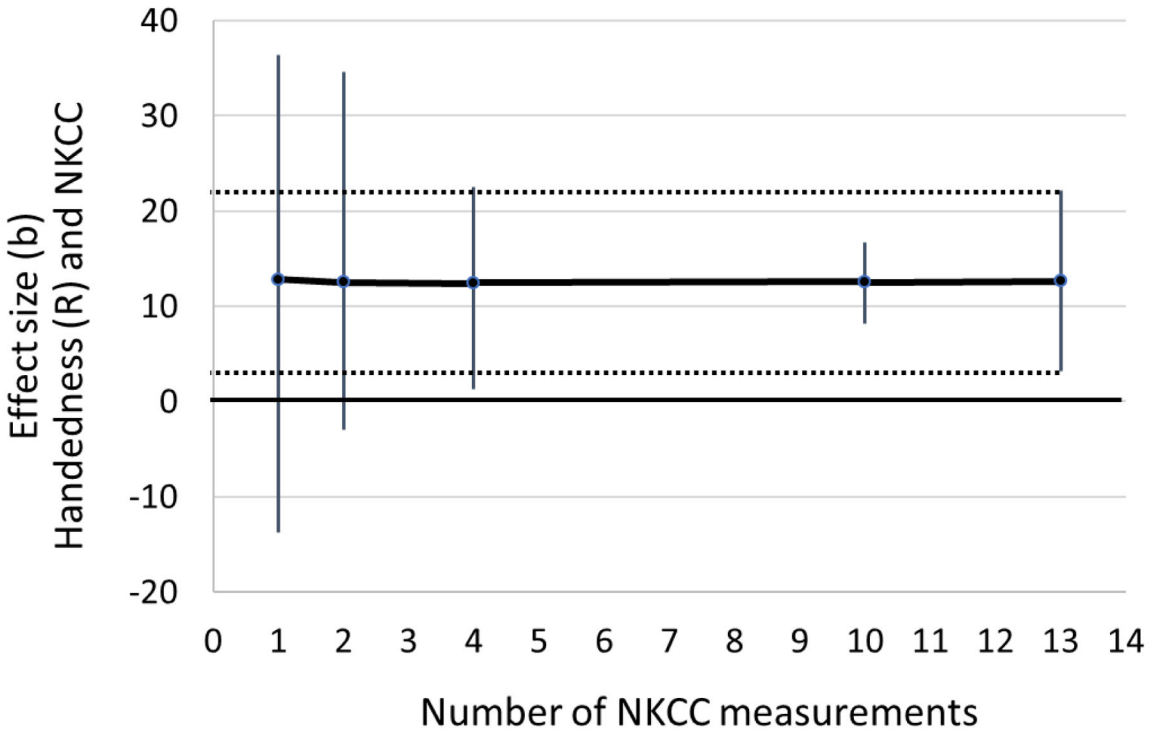

Figure 4.

Comparison of beta weights obtained with averages of 1, 2, 4, and 10 observations (x axis) of natural killer cell cytotoxicity (NKCC) with the best estimate of beta (average of all 13 observations). Sign errors are below the bold line (b = 0), and magnitude errors are outside the dotted lines (the 95% confidence interval of the best beta).

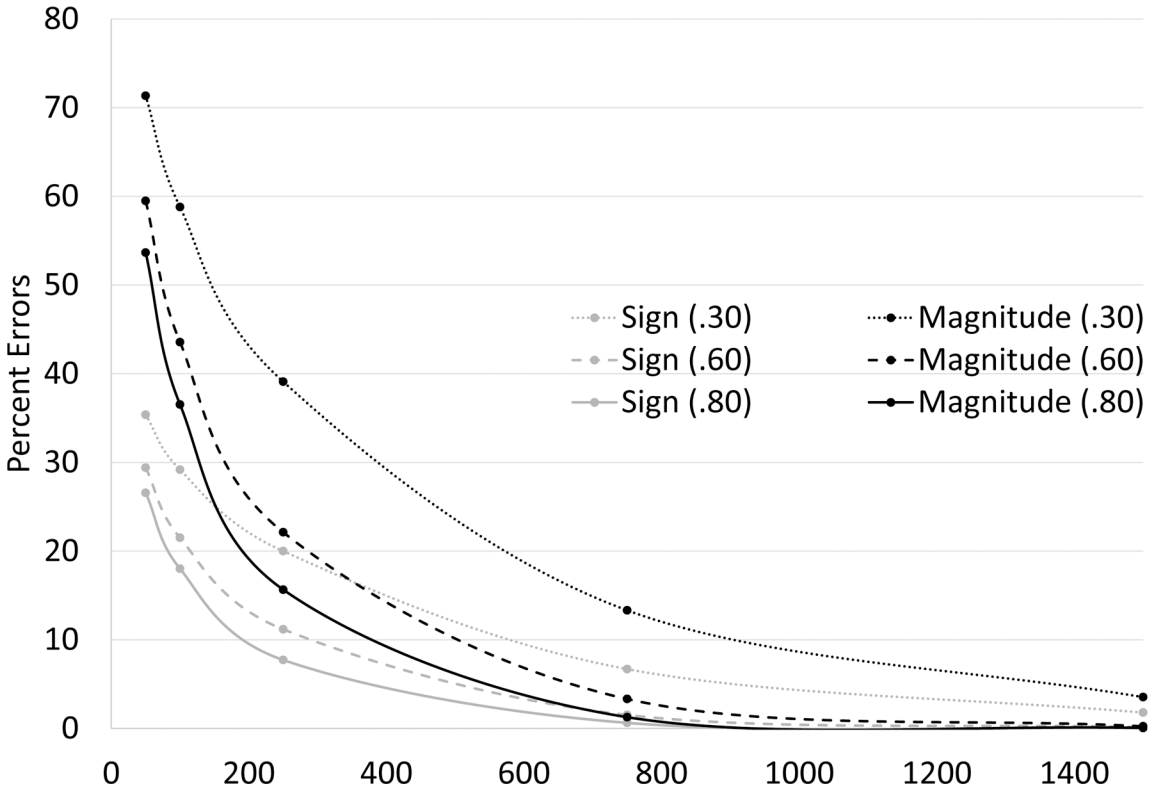

Figure 5.

Results of simulations where the relationship between X and Y is $r = 0.10$ and the reliability of Y is .30, .60, or .80. On the Y axis, percent of sign errors was defined as the proportion of 10,000 simulations in which the beta weight for X was < 0 and magnitude errors, when the obtained beta weight was more than .10 different from the true relationship. On the X axis is sample size in the simulated datasets.[1]

[1]Note the elbow at N ~ 200 in the reliable measurement case. This elbow agrees with the N at which correlation estimates stabilized in another simulation study using different methodology (25).

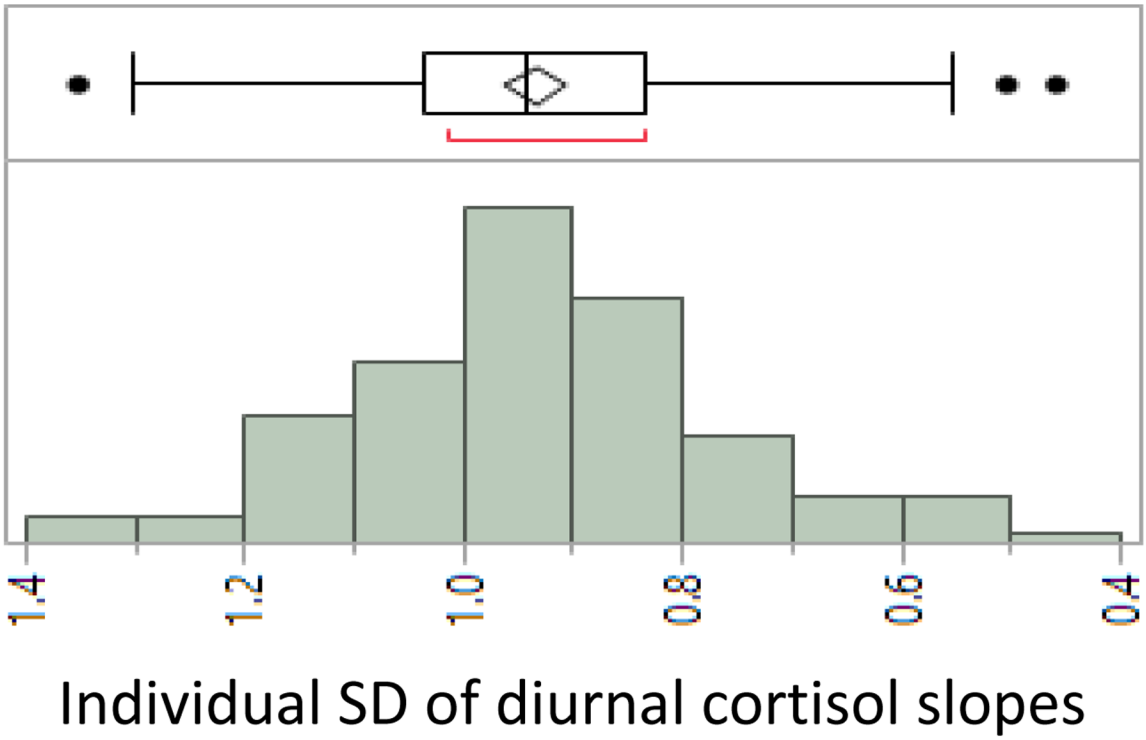

Figure 6.
Histogram and box plot of individuals' standard deviations (iSD) in diurnal cortisol slope among 125 law students measured up to 15 times over 6 months.