volume 16 issue 2
april | 2014

www.ehps.net/ehp

The European Health Psychologist

Bulletin of the European Health Psychology Society

editorial

Health Psychology and statistical methods

Out with the old and in with the new

Marta Marques
co-editor
Kyra Hamilton
co-editor

As psychologists, we have all had exposure to statistics and research methodology. Some of us may have embraced the experience with enthusiasm but others, however, have dreaded and still dread the experience. Despite ones level of enthusiasm or dread, we are still collectively guilty of continuing to use the techniques we were taught in our statistics 101 classes. To continue to conduct rigorous research, we have an obligation to update our knowledge. This Special Issue of the European Health Psychologist Bulletin is dedicated to showcasing a series of papers on a range of statistical approaches that are considered to be more sophisticated and comprehensive alternatives to the methods we currently use.

Highlights of the Special Issue

Three key themes unify the current collection of articles. First, we demonstrate that researchers are applying more sophisticated methods to the analysis and collection of longitudinal data in health psychology research. Second, we highlight the adoption of alternative and more comprehensive assessment methods to assess reliability and validity, a central issue in the measurement of psychological constructs. Finally, we present an article that outlines the application of Bayesian Statistics, an approach that is not necessarily new but often engenders in people some uncertainty towards its use.

In addressing the first theme, Ntoumanis (2014) provides an introduction to a robust and currently recommended statistical approach to analyse longitudinal repeated measures data with a hierarchical structure, Multilevel Modelling (MLM) (see Singer & Willett, 2003). Although MLM has clear advantages over other methods, it is not yet widely used in health psychology research. In this paper, the author presents different methods and procedures employed in MLM, unique advantages for its use, and examples of its application in health psychology research (e.g. motivation for physical activity). Next, Richardson and Fuller-Tyszkiewicz (2014) demonstrate how MLM can be used to analyse intensive longitudinal data collected by means of experience sampling method (ESM; or Ecological Momentary Assessment; Bolger & Laurenceau, 2013). ESM has the advantage of capturing real time emotions, thoughts, and behaviours. Although this approach is increasing in the empirical

1 For further discussion on the use of intensive longitudinal methods also see Stadler et al, also published in the *EHP* (2013, September issue; http://www.ehps.net/ehp/issues/2013/v15iss3_September2013/EHP_September_2013.pdf)

literature, research on the statistical analyses that best model the data obtained with ESM is limited. Using an example of an analysis undertaken to assess the relationship between positive affect and drinking behaviour, the authors illustrate and compare the application of different modelling approaches (log-linear model vs. non-linear models), and present the advantages of using the non-linear threshold dose-response approach to analyse data collected from ESM. In a final paper, Mohr (2014) provides further insight into the various applications of combining ESM and MLM in health psychology research, and explores the potential of using within-person processes (captured with ESM) as predictors of longer-term health-related outcomes, the so-called slopes-as-predictors method (see Mohr et al., 2013).

Our second theme focuses on reliability and validity. Gjalt-Jorn Peters provides a solid argument for abandoning the use of Cronbach's alpha (Cronbach, 1951) as an indicator of the internal consistency of a scale because, as he and others suggest (e.g., Sijstma, 2009) it is unrelated to internal consistency. Several alternative estimates (e.g. Greatest Lower Bound; Sijstma, 2009) have been proposed in the recent literature but, as the saying goes, "old habits are hard to break". To facilitate a transition to the use of these optimal estimates, the author discusses the creation of a user-friendly function to compute these indices in the open source statistical package R that does not necessarily require comprehensive knowledge of this software. In this paper, clear guidelines on how to estimate these alternative measures of reliability are provided, and considerations on the dynamics of reliability and validity and their distinction are discussed. Peters leaves us with the message that the use of a combination of reliability and validity diagnostics to assess scale quality is essential. In a follow-up commentary to the Peters article, Crutzen (2014) proposes that test-retest reliability, an important component of reliability, should also be included in a comprehensive assessment of scale quality. The author argues for the advantages of doing test-retest analysis and discusses available estimates that take into account changes in measurement error due to time (e.g. Coefficient of equivalence and stabilityl; Schmidt, Le, & Ilies, 2003). The author presents the procedures to compute the test-retest estimations in the R package.

In our final theme, the application of Bayesian Statistics is explored (see Kaplan & Depaoli, 2013). Van de Schoot and Depaoli (2014) acknowledge that although most of us have heard or read a few things about this type of statistical procedure, many of us are still clueless about its use, whether we should use it, and how to begin using it. The authors advocate that all types of conventional questions can be addressed with Bayesian statistics. In their article, they provide readers with an introduction to Bayesian statistics and definitions of the key concepts, as well as the advantages of its use over conventional statistical methods. In addition, the authors provide guidelines on how to report the implementation and results of Bayesian methods in empirical articles.

Conclusions

Psychology researchers are required to have a relatively extensive knowledge of statistical methodology and remain up-to-date with novel statistical methods, procedures, and software. The six contributions published in this Special Issue reflect diverse and stimulating perspectives on innovative, alternative and/or increasingly

popular statistical approaches. The collection of papers is intended to be of interest for readers with varying levels of statistical knowledge. It is hoped that these papers will spark both consideration of and/or further debate in the use of statistical methods used in health psychology research.

References

Bolger, N., & Laurenceau, J-P. (2013). *Intensive longitudinal methods: an introduction to diary and experience sampling research*. New York: Guilford.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. doi:10.1007/BF02310555

Crutzen, R. (2014). Time is a jailer: what do alpha and its alternatives tell us about reliability? *The European Health Psychologist, 16*(2), 70-74.

Kaplan, D., & Depaoli. S. (2013). Bayesian statistical methods. In T. D. Little (Ed.), *Oxford handbook of quantitative methods* (pp. 407-437). Oxford: Oxford University Press.

Mohr, C. (2014). Within-person indicators of health. *The European Health Psychologist, 16*(2), 52-55.

Mohr, C., Brannan, D., Wendt, S., Jacobs, L., Wright, R., & Wang, M. (2013). Daily mood-drinking slopes as predictors: a new take on drinking motives and related outcomes. *Psychology of Addictive Behaviors, 27*(4), 944-955. doi:10.1037/a0032633

Ntoumanis, N. (2014). Analysing longitudinal data with multilevel modelling. *The European Health Psychologist, 16*(2), 40-45.

Peters, G-J. Y. (2014). The alpha and the omega of scale reliability and validity: why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *The European Health Psychologist, 16*(2), 56-69.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: an empirical examination of the effects of different sources of measurement error on realibility estimates for measures of individual differences constructs. *Psychological Methods, 8*(2), 206-224. doi:10.1037/1082-989X.8.2.206

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's

Alpha. *Psychometrika, 74*(1), 107–120. doi:10.1007/s11336-008-9101-0

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: modelling change and event occurnace*. New York: Oxford University Press

Richardson, B., & Fuller-Tyszkiewicz, M. (2014). The application of non-linear multilevel models to experience sampling data. *The European Health Psychologist*, 16(2), 46-51.

van de Schoot, R., & Depaoli, S. (2014). Bayesian analyses: where to start and what to report. *The European Health Psychologist, 16*(2), 75-84.

Marta Marques
co-editor of The European Health Psychologist
CIPER, Faculty of Human Kinetics, University of Lisbon, Portugal
martamarques@fmh.ulisboa.pt

Kyra Hamilton
co-editor of The European Health Psychologist
School of Applied Psychology, Griffith University, Australia
kyra.hamilton@griffith.edu.au

original article

Analysing longitudinal data with multilevel modelling

Nikos Ntoumanis
Curtin University

Recent advances in statistical analysis have resulted in the decline of the use of repeated measures ANOVA/MANOVA for the analysis of longitudinal data in health psychology research. One of the more sophisticated and comprehensive alternatives to these tests (see Kwok et al., 2008) is multilevel modelling (MLM), also known as hierarchical linear modelling or linear mixed modelling. MLM is appropriate for the analysis of data with a nested structure, for example, patients (level 1) nested within clinics (level 2). Ignoring the nested structure of such data can result in biased estimates of standard errors and subsequent increase in Type I error (Hox, 2010). MLM is also useful for testing the interaction between individual and contextual factors and exploring heterogeneity in the data due to their nested structure. Many applications of MLM in the health psychology literature incorporate two or three levels of analysis. For example, Mayberry, Espelage and Koenig (2009) examined adolescents' perceptions of parental and peer influence (level 1) and school characteristics (level 2) as predictors of adolescent substance use. In addition to analysing cross-sectional data, MLM can also be used for longitudinal data, given that multiple measurement points (level 1) are nested within individuals (level 2) who can also be nested within a group setting (level 3). For example, Ntoumanis, Taylor and Thogersen-Ntoumani (2012) examined moral attitudes, emotional well-being, and indices of behavioural investment in a sample of British adolescent athletes. In this study, each variable was measured at three time points during a sport season. The three time points were the first level of the analysis, with athletes and their teams constituting the second and third levels of the analysis, respectively.

In this paper, I offer a very brief overview of how multilevel modelling can be employed for the analysis of longitudinal data without presenting any mathematical formulas. I use an example from the physical activity literature to demonstrate, step-by-step, decisions that need to be made with regard to the analysis of the data. I refer the reader to Singer and Willett's (2003) book for a far more detailed treatment of MLM for longitudinal data analysis, including testing the assumptions that underlie such analysis.

MLM can be used when all individuals are assessed on the same number of occasions which are equally spaced over time. However, MLM can also be used when the spacing of measurement points is not identical across individuals (e.g., the time interval between cancer screenings might vary across participants), and also when the number of measurement waves is not the same across individuals. The latter is a particularly important feature, given the attrition of participants recorded in longitudinal studies. As Singer and Willett (2003) note, each individual's growth record can contain a unique number of waves collected at unique occasions of measurement. The impact of missing data on MLM estimates is discussed by Hox (2010).

In its simplest form, a MLM of change is a linear growth model with a random intercept, as well as a random slope to represent change over

time in the dependent variable (note that a model with no growth term can also be calculated initially in order to estimate the intraclass-correlation coefficient which quantifies the variation in the dependent variable across the different levels of the analysis). For example, Figure 1 demonstrates changes in intrinsic motivation (measured on a 1-7 point scale with higher scores indicating greater motivation) for physical activity in children over 6 time points in a three-year period. It is clear from the figure that there was considerable inter-individual variation in the motivation scores at the beginning of the study and in the trajectory of change of motivation over time. Such heterogeneity in the intercept and the rate of change can be captured by including additional predictors in the model. The intercept and growth across the whole sample are shown with the thick dotted line. Such variations cannot be captured in a fixed effects ANOVA model, but can be important from an applied and conceptual perspective. Multilevel modelling provides a statistical test of the variation in both the intercept and the growth terms across individuals (see Model 1 in Table 1).

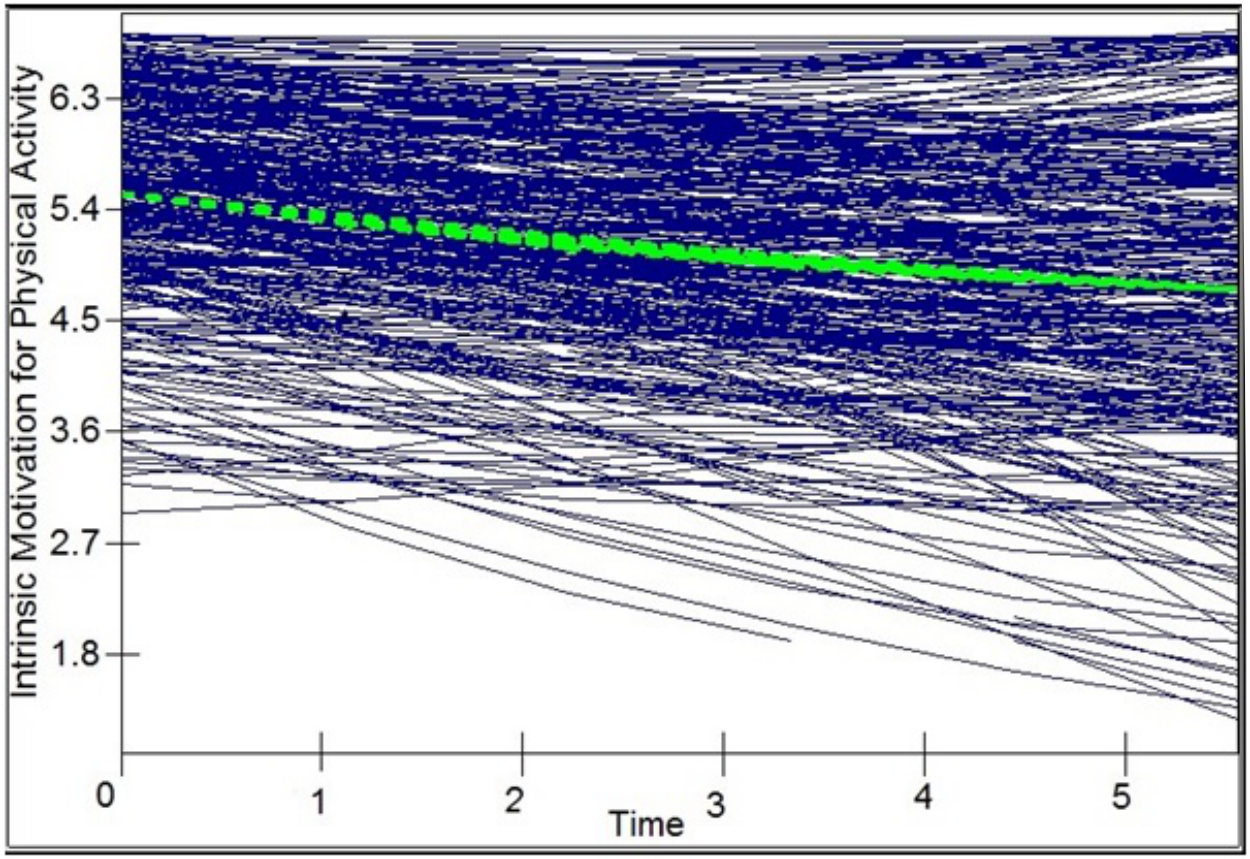


Figure 1: Variation in intercepts and slopes of change in intrinsic motivation for physical activity over six time points

One of the first issues to think about when using MLM to analyse longitudinal data is what type of change to examine. In our example, up to three growth terms can be tested: linear, quadratic and cubic. A study will need at least 3, 4 and 5 time points to test interindividual variation in linear, quadratic, and cubic growth, respectively. Figure 1 includes both linear ($b = -.25$, $p < .01$) and quadratic ($b = .02$, $p = .03$)

Table 1. *Model Examining Changes in Intrinsic Motivation for Physical Activity Over a 3-Year Period*

	Model 1 *B (SE)*	Model 2 *B (SE)*	Model 2 *B (SE)*
Level 1 Predictors			
Intercept	5.526 (.066)	5.529 (0.60)	4.292 (3.15)
Time (linear)	-.254 (.046)	-.249 (.043)	-.249 (.043)
Time (quadratic)	.020 (.009)	.020 (.008)	.020 (.008)
Competence		.320 (.051)	.234 (.055)
Competence x Time (linear)		.071 (.040)	.090 (.040)
Competence x Time (quadratic)		-.008 (0.008)	-.011 (.008)
Level 2 Predictors			
Mean Competence			.249 (.062)
Level 1 Variance	.656 (.033)	.524 (.028)	.521 (.028)
Level 2 Variance			
Intercept	.945 (.119)	.703 (.100)	.694 (.099)
Time (linear)	.155 (.062)	.153 (.055)	.156 (.055)
Time (quadratic)	.004 (.002)	.005 (.002)	.005 (.002)
Competence		.116 (.026)	.112 (.026)
Model Deviance	5507.596	5021.247	5185.553

Note: When the ratio of B/SE is above 1.96, the parameter is considered significant at $p < .05$; this ratio is a z statistic

growth; the cubic growth term was almost zero and was excluded from the equation. Both linear and quadratic terms were significant and their inter-individual variation was also significant (see Model 1 in Table 1). Although not shown in Table 1, it is also useful to inspect covariances between the intercept and the growth terms in order to determine whether participants' initial mean score of intrinsic motivation is related to the rate of change of their scores over time. However, note that it is not always necessary that the intercept represents initial status. In multilevel growth models the growth term or terms for time can be centred by assigning the value of zero to different time points, such as the beginning, middle or end of the study (or any time point of interest), depending on the substantive question pursued in the study. In Figure 1, time is centred (time1 = 0) at the first wave of measurement, hence the intercept of the growth model can be interpreted as students' reports of intrinsic motivation at beginning of the study.

Another issue to consider when analysing longitudinal data with MLM is the type of predictors that can be included in the analysis. In addition to the intercept and the growth terms, additional predictor variables can be added in the multilevel regression equation at different levels of the analysis. In a two-level model (repeated measures nested within individuals) both time-varying covariates (level 1) and time-invariant covariates (level 2) can be introduced. By adding predictors the unaccounted variance at the corresponding level of each predictor can be reduced, however, the unaccounted variance at the other level might either decrease or increase. Singer and Willett (2003) discuss this problem and suggest suitable pseudo-R2 indices.

An example of a level 1 covariate in our example could be perceptions of physical activity competence measured across all six time points. An example of a level 2 covariate could be a personality or demographic variable measured at one point in time. Furthermore, interactions can be tested between predictor variables within the same level or at different levels. In our example, we entered in the multilevel regression equation the additional predictors of perceptions of physical activity competence, as well as the interactions between competence and linear growth and between competence and quadratic growth (see Model 2 in Table 1). Perceptions of competence emerged as significant predictors of intrinsic motivation ($b = .32$, $p < 0.01$) but the two interaction terms were not significant. The effect of each level 2 predictor can be tested as fixed or random. In our example, the main effect of competence could, depending on available theory or evidence, be conceptualised as being the same across all individuals and therefore treated as fixed, or varying from individual to individual and hence treated as random. Whilst treating the slopes of level 2 predictors as random helps researchers answer interesting research questions associated with between-person variability, a model with many random effects might not converge. Singer and Willett (2003) and Hox (2010) offer some detailed guidance for model building and model comparison, involving the inspection of deviance statistics for each model.

In growth models the slope of a level 1 (time-varying) predictor confounds inter-individual change and between-person variability. Hence, it is suggested that the aggregate of each level 1 predictor is entered at the level 2 of the analysis. In our example, if we average perceptions of competence within each individual across all measurement waves, this variable could be entered as a level 2 predictor in the analysis. In the new model (Model 3, Table 1), the slope of the level 1 measure of competence ($b = .23$, $p < 0.01$) represents the within-person association between competence and intrinsic motivation over time, after partialling out between-person differences in competence. However, the two slope terms for competence at the two levels of the analysis might or might not be correlated, depending on how the level 1 predictor has been centred. The issue of centring is often discussed in the MLM literature and is another important factor to consider with this type of analysis. Centring helps the interpretation of their intercepts and slopes but the type of centring has often puzzled researchers unfamiliar with the complexities of the analysis. Enders and Tofighi (2007) and Lüdtke, Robitzsch, Trautwein and Kunter (2009) offer some excellent guidance on centring for cross-sectional multilevel data, and their recommendations also apply for longitudinal MLM data.

Briefly stated, the level 1 predictor scores could be centred around each person's unique mean score over time (group-mean centring; CWC) or across all individuals' mean score over time (grand-mean centring; CGM). In both cases, the level 1 slope is the same, but the level 2 slope will differ. With CGM, the level 1 and 2 slopes are correlated, hence the level 2 slope is a partial effect controlling for level 1. With CWC, the two slopes are uncorrelated, hence the level 2 effect is a mixture of level 1 and level 2 effects (this is the case for the level 2 slope for competence shown in Model 3, Table 1). To obtain a pure estimate of level 2 effect, we need to calculate the difference between the level 2 and level 1 slopes (Marsh et al., 2012); in our example, .249- .234= .015. In brief, if the within-person associations (level 1) are of interest, then either type of centring will provide the same estimate which is not confounded by inter-individual differences. However, if inter-individual differences are of

interest, then the type of centring used will result in different slope estimates. In most cases, however, researchers are interested in Level 1 associations.

Testing linear or non-linear terms for time in a MLM equation is a sensible option when certain trends are expected over time. In other studies (e.g., diary studies) such trends might not be expected. For example, if one is interested in examining dietary lapses over a typical 7-day period, there is no rationale to expect a particular pattern of growth over that period. However, other contrasts of interest could be entered to detect specific trends. For example, McKee, Ntoumanis and Taylor (in press) showed that dietary lapse occurrences were more likely in the evening compared to the morning (b=0.37, p=0.002) and afternoon (b=0.24, p=0.01) over a 7-day period.

Often in health psychology researchers are interested in several dependent variables. In such cases a multivariate growth model can be used instead of several univariate growth models. Specialised MLM software such as MLwiN can perform this analysis by adding one extra level. Other software with structural equation modelling capabilities (e.g., Mplus) can also perform multivariate MLM but with a different set up; in fact, in Mplus the number of levels is one less than the number of levels in conventional MLM software (Muthén & Muthén, 1998-2012). Such software can also perform multilevel structural equation modelling which, unlike standard applications of MLM regressions, take into account measurement error and can test both simple and complex mediation effects (Preacher, Zyphur, & Zhang, 2010).

An often asked question revolves around the sample size needed to perform MLM analysis of change. Various rules of thumb have been proposed in the literature; for example, a simulation study by Maas and Hox (2008) suggests that sample sizes of 50 or less at level 2 can result in biased estimates of the standard error of the variance terms in that level. The regression coefficients and level 1 variance terms are fortunately not affected by this bias. Maas and Hox's simulation had 5 observations as the minimum number at level 1 (in other words, number of repeated measures for each individual in a longitudinal MLM). A better option than rules of thumb and simulation studies is the use of specialised software to calculate the sample size requirements for a particular study. A freely available software for power analysis, for both cross-sectional and longitudinal MLM is Optimal Design, available at http://sitemaker.umich.edu/group-based/optimal_design_software . For a two-level longitudinal MLM, the software requires input of values regarding the duration of the study, the frequency of observations, the level 1 variance, the between-person variability in the parameter of interest, and an estimate of effect size. It is also important that researchers build in estimates of expected attrition rates in their calculations.

In sum, MLM can address all the research questions that repeated measures ANOVA/MANOVA tests address without being constrained by the rigid assumptions of the latter (see Kwok et al., 2008). Further, MLM can be used to pursue research questions that cannot be answered with repeated measures ANOVA/MANOVA. Health psychologists can benefit in many ways from using MLM in their analysis of longitudinal data. Many commercial (etc., MLwiN, HLM, Mplus, SPSS, SAS) and some free software (R) can be used for such analysis.

References

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue.

Psychological Methods, 12(2), 121-138. doi:10.1037/1082-989X.12.2.121

Hox, J. J., (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Kwok, O. M., Underhill, A. T., Berry, J. W., Luo, W., Elliott, T. R., & Yoon, M. (2008). Analyzing longitudinal data with multilevel models: An example with individuals living with lower extremity intra-articular fractures. *Rehabilitation Psychology, 53*(3), 370-386. doi:10.1037/a0012765

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*(2), 120-131. doi:10.1016/j.cedpsych.2008.12.001

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*(3), 86-92. doi:10.1027/1614-2241.1.3.86

McKee, H., Ntoumanis, N., Taylor, I.M. (in press). An ecological momentary assessment of lapse occurrence in dieters. *Annals of Behavioral Medicine*. doi: 10.1007/s12160-014-9594-y

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*(2), 106-124. doi:10.1080/00461520.2012.670488

Mayberry, M. L., Espelage, D. L., & Koenig, B. (2009). Multilevel modeling of direct effects and interactions of peers, parents, school, and community influences on adolescent substance use. *Journal of Youth and Adolescence, 38*(8), 1038-1049. doi:10.1007/s10964-009-9425-9

Muthén, L.K. and Muthén, B.O. (1998-2012). *Mplus User's Guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Ntoumanis, N., Taylor, I. M., & Thøgersen-Ntoumani, C. (2012). A longitudinal examination of coach and peer motivational climates in youth sport: Implications for moral attitudes, well being, and behavioral investment. *Developmental Psychology, 48*(1), 213. doi:10.1037/a0024934

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15*(3), 209-233. doi:10.1037/a0020141

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Nikos Ntoumanis
is Professor of Psychology at the Centre of Health Psychology and Behavioural Medicine, School of Psychology and Speech Pathology, Curtin University, Australia

nntoumanis@gmail.com

original article

The application of non-linear multilevel models to experience sampling data

Ben Richardson
Deakin University
Matt Fuller-Tyszkiewicz
Deakin University

A great deal of evidence demonstrates that state-based aspects of human functioning, such as moment-to-moment variation in affect, explain important psychological and behavioural outcomes (e.g., Colautti et al., 2011); often over and above more general measures that may be used in cross-sectional designs (Sturgeon & Zautra, 2013). For example, in clinical samples, findings demonstrate that a common feature of many disorders is higher levels of reactivity following stressful events (MyinGermeys et al., 2009).

State-based aspects of thoughts and behaviours can be assessed using the experience sampling method (ESM, aka ecological momentary assessment) (Bolger & Laurenceau, 2013; Csikszentmihalyi & Larson, 1987). ESM is a form of intensive longitudinal data collection where participants repeatedly respond, commonly multiple times per day, to questionnaires that assess their experience "right now". Participants' responses may be cued by prompts that occur at random or fixed intervals or by an event (e.g., when the person exercises). Although the method may be burdensome for participants and researchers (Palmier-Claus, MyinGermeys, Barkus, Bentley, & Udachina, 2011), the observations obtained have the advantages of offering a precise test of temporal relationships between variables of interest and increased ecological validity (Bolger & Laurenceau, 2013).

ESM data collection yields a hierarchical dataset where a series of observations (i.e., single responses at a particular time point) are nested within participants. A range of modeling options exists to analyse nested data including regression with robust standard errors and multilevel modeling (MLM). Although MLM is more complex than traditional regression, it allows explicit investigation of individual variability in relationships (i.e., investigation of 'random effects'). For example, a traditional regression approach to studying the relationship between affect in the morning and subsequent drinking in the evening assumes that this relationship is constant across individuals. However, it is possible that some individuals' drinking is more influenced by their mood than others; in other words, that the relationship between affect and subsequent drinking will be stronger for those particular individuals. MLM can test this possibility and also explain variation in this relationship using individual level variables (i.e., an individual's trait coping or impulsivity could explain variation in the strength of the relationship). For this reason, MLM is commonly employed to analyse ESM data.

Within the MLM framework, most commonly relationships between variables are represented in a linear fashion using either linear regression (continuous DV) or logistic regression (binary DV). Although in many cases a linear model may accurately represent the data, it is not guaranteed that the relationships are linear and other relationships are possible. Given this, when analysing ESM data, we recommend undertaking a comprehensive strategy that investigates a range of possible relationships.

More accurate modeling of relationships will contribute to greater understanding of the phenomena of interest.

In this paper we demonstrate such an approach in the context of an analysis undertaken to assess the relationship between positive affect and risky single occasion drinking (RSOD; consumption of 5+ standard drinks in one sitting). The study involved 37 participants (8 males; 29 females) responding to a smartphone-based questionnaire four times per day for ten days. At each time-point, the questionnaire measured participants' mood and whether they had engaged in RSOD. A baseline questionnaire included measures of demographic information and impulsivity (fun seeking and drive). In the context of this dataset, three models are illustrated and compared: a traditional linear model and two alternative models useful for studying non-linear effects: a piecewise regression model and a threshold dose-response model (Hunt & Rai, 2003).

Statistical Models

Traditional model

Commonly, ESM data are analysed using a log-linear model (a multilevel logistic regression) (Hox, 2002). In this model, a binary dependent variable (e.g., RSOD) is regressed onto one or more independent variables (e.g., previous positive mood). This is represented below in equation 1, where i represents the i^{th} individual and j represents the j^{th} assessment point; β_{01i} represents the intercept for the Level 1 equation (i.e., the average probability of engaging in risky drinking); β_{10i} is the unstandardised coefficient representing the relationship between the independent variable and the dependent variable (i.e., the relationship between positive mood and RSOD). β_{01i} is the random effect representing individual differences in the Level 1 IV-DV relationship (i.e., individual differences in the strength of the relationship between positive affect and RSOD). In the event that this random effect is significant, β_{01i} is regressed onto Level 2 (individual difference) variables (in this case: age, gender, fun seeking, drive). This is shown in equation 2, where γ_{001} is the intercept for the Level 2 model; γ_{001} - γ_{013} are the unstandardized coefficients representing the moderating influence of the Level 2 variables on the relationship between positive mood and drinking; u is the error term for Level 2.

Piecewise regression model

This model assumes that there is a cutting point (or knot) on the IV continuum at which the slope of the relationship between IV and DV changes. In a standard piecewise regression, the researcher must pre-specify the value of the knot (i.e., the value where the relationship between positive affect and RSOD changes) in

$$logit(Drink_{ij}) = \beta_{01i} + \beta_{10i} * (positive\,mood) + e_{ij} \quad \text{(Equation1)}$$

$$\beta_{10} = \gamma_{001} + \gamma_{010} * (age) + \gamma_{011} * (gender) + \gamma_{012} * (fun) + \gamma_{013} * (drive) + u \quad \text{(Equation 2)}$$

order to run the model. In the absence of prior evidence for what that cut value should be, researchers may trial different values. In brief, the equation incorporates two key predictors representing the slope below and above the knot. When an individual scores below the knot, the second predictor (above the knot) drops out of the equation:

$$logit(Drink_{ij}) = \beta_{01i} + \beta_{10i} * (positive\,mood) + D * \beta_{11i} * (positive\,mood - t)$$

$$D = \begin{cases} 0 & \text{positive mood } <\text{t} \\ 1 & \text{positive mood } \geq\text{t} \end{cases} \qquad \text{(Equation 3)}$$

Where β_{01i} represents the intercept; β_{10i} represents the slope below the knot; β_{11i} represents the slope at or above the knot; D is the dummy variable representing whether the knot value (t = cutting value on positive mood) has been met/exceeded (D=1) or not (D=0).

Threshold dose-response model

This model is differentiated from the traditional log-linear model in that it includes a threshold value around which the shape of slope for the IV-DV (i.e., positive affect-RSOD) relationship changes, thus in effect producing two lines of best fit (equation 4). The basis for this model is the notion that the relationship between the IV and DV is negligible (~ zero relationship) below a threshold because low-level exposure fails to influence the likelihood of the target event. Once exposure (in our example, positive mood) exceeds this threshold, a positive linear relationship between exposure level and likelihood of outcome (risky drinking) is evident. Importantly, the threshold level is empirically derived from the data rather than needing to be pre-specified by the researcher.

Where logit($drink_{ij}$) is the probability of drinking expressed in logit form; τ is the threshold dose of positive mood; β_0 is the intercept; β_1 is the slope parameter above the threshold; d is the actual dose (i.e., level of positive mood). As implied by Equation 4, the probability of a drinking episode is held constant when positive mood is below the threshold, and exhibits a dose response relationship beyond the threshold (see Figure 1).

$$Logit(Drink_{ij}) = \begin{cases} \beta_0 & \text{f or } d_i < \tau \\ \beta_0 + \beta_1(d_i - \tau) & \text{f or } d_i \geq \tau \end{cases} \qquad \text{(Equation 4)}$$

Data Analytic Strategy

Overview

The utility of three models was explored in the context of the relationship between positive mood and drinking. In each of the models, positive mood at one time point was used to predict likelihood of RSOD (Yes/No) at the next time point, in order to uphold the longitudinal nature of the data and to demonstrate temporal precedence of positive mood. The non-independence of observations arising from the repeated measures design was controlled using

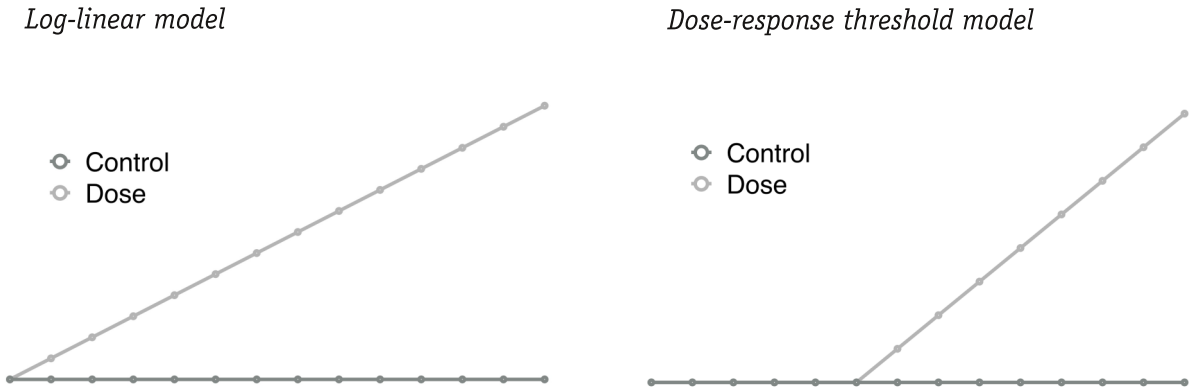


Figure 1: On the left-hand side is a standard log-linear representation of the relationship between positive mood and probability of drinking (traditional model), whereas the panel on the right shows the threshold model, which consists of two separate lines of best fit (a flat line for sub-threshold levels of positive mood, and accelerating probability proportional to exposure beyond the threshold level).

MLM. In each of the three models, random effects were tested for significance to determine whether the strength of the positive mood-drinking relationship varied from individual to individual.

Model comparison

The following indices are used to facilitate comparison of the different modeling approaches: (1) Odds ratios (ORs) were compared in order to compare strength of the IV-DV relationship, (2) standard errors of ORs permitted assessment of precision of these parameter estimates, and (3) log-likelihood, AIC, and BIC values were consulted to make comparisons of fit between these non-nested models, with the lower BIC value having best fit relative to the other models tested. We follow STATA convention of classifying a difference in BIC>10 between two competing models as strong support for the model with the lower BIC value.

Results

The standard multilevel logistic regression suggests that positive mood does not reliably predict the likelihood of a drinking episode (OR = 1.02, se = .012, p = .334). Moreover, this effect failed to vary significantly across individuals (Z = 0.02, p = .986).

The piecewise regression model was fit with different cutting points for the knot (10, 20, 30, and 40), and the best fitting model was achieved when positive mood was split above and below 30. Even so, in this model the positive mood-drinks relationship was positive but non-significant both below the knot (OR = 1.01, se = 0.019, p = .679) and above the knot (OR = 1.02, se = .03, p = .499). Furthermore, the two slopes failed to significantly vary (Z<30 = 0.285, p = .776; Z≥ 30 = 0.227, p = .821).

Finally, the threshold model suggests that the relationship between positive mood and likelihood of drinking is negative below the threshold (OR = 0.97, se = .11, p = .768) and positive above the threshold (OR = 1.01, se = .02, p = .566), but neither effect was significant. However, when these slopes were allowed to vary, the slope above the threshold significantly differed across participants (Z = 9.88, p < .001). Individual differences in this slope were regressed onto key trait-level variables, and it was found that the slope was

strongest for individuals who were older (γ_{010} = .003, p <.001), male (γ_{011} = .009, p = .032), and who reported tendency to engage in behaviors because they are perceived as fun (γ_{012} = .002, p = .014). Reward drive was not a reliable moderator of the positive mood-drinking relationship (γ_{013} = -.001, p = .379). Finally, the slope below the threshold did not differ across individuals (Z = 0.28, p = .779).

Comparison of model fit statistics

As shown in Table 1, the threshold model produced the best fit of the data, followed by the traditional model and then the piecewise model. Using a difference of BIC > 10, the improvement in fit when using the threshold approach relative to the other two approaches provides strong support for this model.

Table 1. *Comparison of fit statistics for the three models*

Model	Log Likelihood	AIC	BIC
Traditional	-272.03	550.11	563.64
Spline	-281.51	570.01	589.17
Threshold	-247.43	504.86	527.39

Note: AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion

Discussion

Despite a significant increase in the volume of experience sampling studies (Mehl & Connor, 2011), there has been limited consideration of how to optimally model the state-based associations captured with this study design. The present study demonstrates several different modeling approaches for their ability to model the relationship between positive affect and likelihood of engaging in RSOD.

Although the positive affect-drinking relationship was weak across each of the tested models, the benefits of a threshold dose-response approach were still evident. First, this threshold model was the only model to detect that the relationship between positive affect and drinking has a negative slope at low levels of positive affect. The traditional multilevel logistic regression approach summarises a single line of best fit, and suggested that the relationship is positive. The piecewise approach also suggested that the relationship is positive across the range of positive affect levels, although the relationship may be slightly stronger at higher levels of positive affect. The stronger performance of the threshold model is further supported by commonly used model fit statistics (log likelihood, AIC, BIC), which suggested that the threshold approach provided a meaningful improvement in correspondence with the data relative to the other two models. Third, the threshold model was the only one to identify random effects for the positive affect-drinking relationship, and these random effects were in turn linked with age, gender and fun-seeking.

A further advantage of the threshold approach over the piecewise approach is that the former empirically derives the appropriate cutting point/threshold, whereas the latter requires researchers to pre-specify the cutting point(s) and then test their plausibility. This pre-specification threatens to be inaccurate: in instances where a predictor with a large range of scores is modeled, there are many different points to be possibly tested, increasing the likelihood that the researcher will miss the appropriate value. Indeed, although we presented results for the best of several knots tested, the poor performance of the piecewise model in this study may derive from choice of knot value.

It should be noted that the added complexity of the piecewise and threshold models appeared to come at a cost to efficiency in estimation as

the standard error for the odds ratio of the slope in the traditional model was lower than the standard errors for any of the parameter estimates in the other two models. This is consistent with previous studies (e.g., Hunt & Rai, 2003). Insofar as this is a common effect in these models, the implication is that power may be lower when using this analysis, relative to a standard logistic regression model, and thus would necessitate a larger sample size and/or routine inclusion of covariates that may serve to reduce error variance.

References

Bolger, N., & Laurenceau, J-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York, NY: Guilford.

Colautti, L., Fuller-Tyszkiewicz, M., Skouteris, H., McCabe, M., Blackburn, S., & Wyett, E. (2011). Accounting for fluctuations in body dissatisfaction. *Body Image, 8*(4), 315-321. doi:10.1016/j.bodyim.2011.07.001

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *The Journal of Nervous and Mental Disease, 175*(9), 526-536. doi:10.1097/00005053-198709000-00004

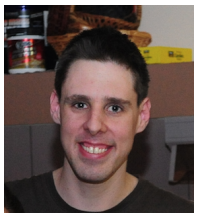
Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Hunt, D., & Rai. S. N. (2003). A Threshold Dose-Response model with random effects in teratological experiments. *Communications in Statistics - Theory and Methods, 32*(7), 1439-1457. doi:10.1081/STA-120021567

Mehl, M. R., & Conner, T. S. (2011). *Handbook of research methods for studying daily life*. New York, NY: Guilford.

MyinGermeys, I., Oorschot, M., Collip, D., Lataster, J., Delespaul, P., & van Os, J. (2009). Experience sampling research in psychopathology: opening the black box of daily life. *Psychological Medicine, 39*(9), 1533-1547. doi:10.1017/S0033291708004947

Palmier-Claus, J.E., MyinGermeys, I., Barkus, E., Bentley, L., & Udachina, A. (2011). Experience sampling research in individuals with mental illness: reflections and guidance. *Acta Psychiatrica Scandinavica, 123*(1), 12-20. doi:10.1111/j.1600-0447.2010.01596.x



Ben Richardson
is Associate Head of School at the School of Psychology at Deakin University, Australia

ben.richardson@deakin.edu.au

Matthew Fuller-Tyszkiewicz
is Senior Lecturer in Psychology (Research Methods) at the School of Psychology at Deakin University, Australia

matthew.fuller-tyszkiewicz@deakin.edu.au

original article

Within-person indicators of health

Cynthia Mohr
Portland State University

With healthcare costs increasing in many nations, a better understanding of risk factors and health treatment effects is needed to enhance health prevention and treatment efficacy. Individuals are not affected similarly by risk factors, nor do they respond uniformly to treatment (overall or day-to-day). Further, some research questions do not lend themselves well to experimentation or traditional longitudinal research (e.g., stress-induced alcohol consumption), thereby complicating measurement. In particular, many variables (e.g., pain, mood, perceived control) are dynamic, with substantial fluctuation (within-person variability), and thus differentially predict outcomes (e.g., Eizenman, Nesselroade, Featherman, & Rowe, 1997). Given the importance of estimating variability in a phenomenon, single time-point measurements are insufficient in assessing the full range of within-person experiences (as opposed to the mean levels). To address these issues, some interventions (e.g., psychotherapy, weight-loss programs) employ journaling or daily surveys to capture within-person reactivity to internal and/or external cues and immediate treatment effects.

Whereas single variable fluctuation is interesting, many research questions concern relationships between multiple fluctuating variables. Researchers have turned to daily process or experience sampling methods combined with multilevel modeling to examine within-person associations among psychosocial variables and health behavior/outcomes, such as one's psychological mood reactivity to stressors or behavioral responses to a positive or negative experience. Results emerging from multilevel modeling analyses, then, involve an intercept and slope of a given person's relationship (such as stress-negative mood), which equates to each individual's own regression equation. The resulting slope provides an estimate of the extent to which people typically respond in a particular way when certain internal or external events occur. Thus, some people may have a more exaggerated or more reactive negative mood response to a given stressor than other individuals, as indicated by significant slope variance. Likewise, some people may have a more positive boost from an intervention activity or stimulus than others. For example, Erica might experience a significant increase in positive mood compared to her typical level of positive mood following supportive interactions. However, Amelia might not experience much change in positive mood, or indeed may actually experience decreases in positive mood. What this approach captures is the dynamic, day-to-day fluctuations that happen within the individual (i.e., some days may be more reactive than others), which is reflected in a positive, negative or neutral slope estimate.

Simultaneously, this approach also measures differences between individuals in that some people may show greater reactivity or responsiveness than others, as depicted in the example with Erica and Amelia. Indeed, this approach is conceptually similar to Mischel and Shoda's (1995) formative work, wherein they define personality as a series of stable but

distinctive if-then situation-behavior signatures, as opposed to previous work conceptualizing individual differences as cross-situationally consistent (Mohr et al., 2013). According to Mischel and Shoda, rather than predicting cross-situational consistency, one should look for reliable patterns of expected behavior within a particular context. So, for example, Amelia might not respond favorably to all socially supportive exchanges, but rather only those that she perceived as helpful or wanted.

Recently, however, researchers have begun to model within-person associations as predictors, rather than outcomes, as they represent potent indices of treatment responsiveness or health-risk reactivity. This addresses a conceptually similar question to one posed by Nesselroade and colleagues, who examined whether intra-individual variability in dynamic and fluctuating factors, (e.g., pain or moods), are powerful predictors of critical health outcomes over and above mean levels, including mortality (e.g., Eizenman et al., 1997). This approach (i.e. slopes-as-predictors) offers a significant advance for health psychologists to predict longer term health and well-being outcomes that contribute above and beyond mean levels of a given variable, such as stress or drinking. It also offers an assessment of the implications of these within-person associations that many of us have been studying for some time; for example, what does it mean that people have a more reactive response to negative events in terms of their well-being over time? What the slopes-as-predictors approach contributes is a unique and more objective measure of how variables that are naturally dynamic and fluctuating relate to longer-term outcomes, compared to more subjective, one-time measures. In particular, information gleaned from within-person associations assessed by repeated measures over time involves contingencies (e.g., stressor-mood; stress-drinking) that are likely outside the awareness of individuals. Indeed, the mechanism by which within-person slopes affect health outcomes is distinct from that by which mean levels influence the same outcomes, akin to the theoretical distinction of stressor exposure and stressor reactivity (Almeida, 2005). Similarly, slopes can predict in the opposite direction from what one might predict based on mean levels, as will be shown below. Although variations exist in this approach, one simple, straightforward method involves extracting individual person-level slopes from a multilevel modeling program, and then employing those as predictors of longer-term outcomes in linear regression equations (while controlling for baseline levels of the outcome; see Mohr et al., 2013).

Much of the existing work using the slopes-as-predictors method has focused on affective reactivity. One set of studies considering these relationships examined within-person negative affect reactivity, measured as same-day and next-day negative affect response to daily stressors, to predict responsiveness to cognitive therapy (Cohen et al., 2008; Gunthert, Cohen, Butler, & Beck, 2005). Results revealed that those who had greater next-day affect spillover responded less quickly to therapy compared to those with lower spillover. Negative event reactivity has also been linked to higher subsequent levels of depression (Parrish, Cohen, & Laurenceau, 2011). Another set of studies demonstrated that those with higher levels of affective reactivity experienced higher levels of general affective distress and likelihood of affective disorder after ten years (Charles, Piazza, Mogle, Sliwinski, & Almeida, 2013), as well as enhanced risk of chronic physical health conditions ten years later (Piazza, Charles, Sliwinski, Mogle, & Almeida, 2012).

My colleagues and I have also recently employed this approach in exploring consequences of behavioral reactivity (i.e. alcohol consumption) to daily positive and

negative mood experiences. In particular, our work has examined outcomes related to within-person mood-drinking relationships in a sample of moderate-to-heavy drinkers (Mohr et al., 2013). We revealed that negative mood-related solitary consumption was associated with lower levels of drinking-to-cope motivations twelve months later. This finding is particularly revealing in that it contradicts the prediction based on mean levels of consumption (i.e., greater consumption predicted stronger motives) and research examining self-reported alcohol use motivations. Although self-reports of drinking as a coping strategy typically predict negative health outcomes, such as alcohol abuse, our assessment of the relationship between negative mood-drinking slopes and follow-up drinking-to-cope motivation indicated a different (and less detrimental) outcome. Our conclusions support that daily mood-drinking associations are a distinct measure from self-reported coping motives. One explanation for our pattern of results may be that, consistent with the work recovery literature (Repetti, 1992), our participants socially withdrew on more stressful days to rejuvenate, which reduced coping motives for drinking over the longer term (at least among moderate drinkers). In contrast, participants who drank more alone on days with increases in positive mood actually demonstrated higher coping motives and lower social motives a year later. Although further research is needed to establish a firm understanding of this result, the positive mood-drinking alone relationship could serve as an index of relationship deficits, whereby these individuals may not have others with whom to share or capitalize on positive experiences (one potential byproduct of social, experience-enhancement drinking). In employing this approach, then, we may have uncovered a new behavioral risk factor for subsequent health problems, such that consistently drinking alone following increases in positive moods is consequential to health over time. Thus, we conclude that how and when people consume alcohol may be at least as important as how much they consume -information that cannot be gleaned from traditional self-report/survey methodology.

In sum, the slopes-as-predictors approach holds much promise for health psychologists striving to gain a better understanding of the interrelationships between psychosocial factors and health outcomes over time. It also affords a new tool for psychologists already interested in dynamic and fluctuating phenomena measured as within-person associations in their short-term context, in relation to longer-term outcomes. Lastly, the benefit of considering individual differences in within-person reactivity processes facilitates better prediction of longer-term health and well-being outcomes, which ultimately should improve prevention efforts.

Author's Note

Funding was provided by NIAAA grants R03-AA014598 and R29AA09917, Faculty Enhancement and SRI support from Portland State University.

References

Almeida, D.M. (2005). Resilience and vulnerability to daily stressors assessed via diary methods. *Current Directions in Psychological Science, 14*(2), 64–68. doi:10.1111/j.0963-7214.2005.00336.x

Charles, S. T., Piazza, J. R., Mogle, J., Sliwinski, M. J., & Almeida, D. M. (2013). The wear and tear of daily stressors on mental health. *Psychological Science, 24*(5), 733–741. doi:10.1177/0956797612462222

Cohen, L. H., Gunthert, K. C., Butler, A.C., Parrish, B. P., Wenze, S. J., & Beck, J. S. (2008). Negative affective spillover from daily events predicts early response to cognitive therapy for depression. *Journal of Consulting & Clinical Psychology, 76*(6), 955–965. doi:10.1037/a0014131.

Eizenman, D. R., Nessleroade, J. R., Featherman, D. L., & Rowe, J. W. (1997). Intraindividual variability in perceived control in an older sample: The MacArthur successful aging studies. *Psychology and Aging, 12*(3), 489-502. doi:10.1037/0882-7974.12.3.489

Gunthert, K., Cohen, L., Butler, A., & Beck, J. (2005). Predictive role of daily coping and affective reactivity in cognitive therapy outcome: Application of a daily process design to psychotherapy research. *Behavior Therapy, 36*(1), 79–90. doi: 10.1016/S0005-7894(05)80056-5

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*(2), 246-268. doi:10.1037/0033-295X.102.2.246

Mohr, C. D., Brannan, D., Wendt, S., Jacobs, L., Wright, R. & Wang, M. (2013). Daily mood-drinking slopes as predictors: A new take on drinking motives and related outcomes. *Psychology of Addictive Behaviors, 27*(4), 944-955. doi:10.1037/a0032633

Parrish, B. P., Cohen, L. H., & Laurenceau, J.-P.(2011). Prospective relationship between negative affective reactivity to daily stress and depressive symptoms.*Journal of Social and Clinical Psychology, 30*(3), 270–296. doi:10.1521/jscp.2011.30.3.270.

Piazza, J. R., Charles, S. T., Sliwinski, M. J., Mogle, J., & Almeida, D. M. (2012).Affective reactivity to daily stressors and long-term risk of reporting a chronic physical health condition.*Annals of Behavioral Medicine, 45*(1), 110–120. doi:10.1007/s12160-012-9423-0

Repetti, R. L. (1992). Social withdrawal as a short term coping response to daily stressors. In H. S. Friedman (Ed.), *Understanding Hostility, Coping, and Health* (pp. 151-161). Washington, DC, US: American Psychological Association.

Cynthia Mohr
is Associate Professor of Applied Social Psychology, Department of Psychology, Portland State University, USA
cdmohr@pdx.edu

original article

The alpha and the omega of scale reliability and validity

Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality

Gjalt-Jorn Y. Peters
Open University of the Netherlands

Health Psychologists using questionnaires rely heavily on Cronbach's alpha as indicator of scale reliability and internal consistency. Cronbach's alpha is often viewed as some kind of quality label: high values certify scale quality, low values prompt removal of one or several items. Unfortunately, this approach suffers two fundamental problems. First, Cronbach's alpha is both unrelated to a scale's internal consistency and a fatally flawed estimate of its reliability. Second, the approach itself assumes that scale items are repeated measurements, an assumption that is often violated and rarely desirable. The problems with Cronbach's alpha are easily solved by computing readily available alternatives, such as the Greatest Lower Bound or Omega. Solving the second problem, however, is less straightforward. This requires forgoing the appealing comfort of a quantitative, seemingly objective indicator of scale quality altogether, instead acknowledging the dynamics of reliability and validity and the distinction between scales and indices. In this contribution, I will explore these issues, and provide recommendations for scale inspection that takes these dynamics and this distinction into account.

Psychologists do not have it easy. Whereas researchers in chemistry, medicine, or physics can usually directly observe the objects of their study, researchers in psychology not only have to rely on indirect measurement of the variables of interest, but these measurements are also subject to a plethora of biases and processing quirks that are not yet fully understood. Whereas biological measures, using for example electroencephalograms or functional magnetic resonance imaging, provide direct access to what are generally considered proxies of psychological activity, most psychologists are limited to measuring behavior. Although behavior is sometimes the variable of interest itself, psychologists often use participants' behavior to measure psychological variables. For example, implicit association tasks present participants with various stimuli and measure how fast participants respond to different stimuli, with the aim of inferring how strongly hypothesized psychological variables are associated; and questionnaires present participants with various items and measure which answer options participants endorse, with the aim of inferring the value of hypothesized psychological variables.

The indirect nature of these measurements leaves much room for unknown sources of variance to contribute to participants' scores, which translates to a relatively low signal to noise ratio, or a proportionally large measurement error. This is detrimental to studies' power to draw conclusions as to associations between the variables under investigation. To ameliorate this situation, researchers often use multiple measurements that are then aggregated. This process decreases the error variance, because as the number of aggregated measurements increases, those parts of the error variance that are not systematic cancel each other out more and more (since, conveniently, researchers usually assume that error variance is random). Of course, this

approach requires *repeated* measurements; if a researcher devised three additional questionnaire items to strengthen the measurement of the construct tapped by a first original item, the three additional items must measure the same construct as the first item. If they measure something else instead, they will decrease the validity of the measurement by adding a source of systematic measurement error. Thus, because psychologists are condemned to indirect measurements of psychological variables, aggregating our measurements is a valuable instrument; but at the same time, caution is advised when aggregating separate measurements into a scale.

Most researchers understand this, and perhaps this is one reason why researchers routinely report Cronbach's Alpha, which is widely considered almost as a quality label for aggregate variables. Researchers and reviewers alike are satisfied by high values of Cronbach's Alpha (many researchers will cite a value of .8 or higher as acceptable), and in fact, inter-relations of items are rarely inspected more closely if Cronbach's Alpha is sufficiently high. This reliance on Cronbach's alpha is unfortunate, yet has proven quite hard to correct (Sijtsma, 2009). One of the reasons may be a combination of self-efficacy and a lack of clear guidelines. Articles addressing the problems with Cronbach's Alpha tend to be quite technical, and rarely provide a tutorial as to what to do instead of reporting Cronbach's Alpha (Dunn, Baguley, & Brunsden, 2013, being a notable exception). The current paper aims facilitate improved scale scrutiny by doing three things. First, a brief non-technical explanation is provided as to why Cronbach's Alpha should be abandoned. Second, alternatives are introduced that are easily accessible with user friendly, free tools, and a tutorial of how to compute these alternatives is provided. Third, a plea is made to step away from convenient quantitative measures as means of assessing scale quality.

Why abandon Cronbach's Alpha

Imagine that we want to measure 'connectedness with the European Healthy Psychology Society (EHPS)' with four items. Figure 1 shows these four items in the simplest possible situation: they are all exactly the same. Of course, this never happens; and Figure 2 shows a more realistic picture. The gray normal curves in the background depict the population distributions for each item. In addition, for each item, the scores of three individuals are shown. When an individual answers each item, each single measurement, depicted by a black dot, is determined by the individual's true score on that item, represented by vertical dotted lines, and measurement error, represented by normal curves that show the likelihood of obtaining given measurements. In Figure 2, "How do you feel about the EHPS?" has considerably more measurement error than "How many EHPS conferences have you attended?". This might be, for example, because factors such as mood and whether somebody happens to have just gotten a submission to Psychology & Health accepted or rejected are more likely to temporarily influence somebody's appreciation of the EHPS than their recollection of the number of attended EHPS conferences. Another difference between the items in Figure 2 are the means: for example, naturally the mean for "How often do you read the EHP?" is exceptionally high. Finally, the variance in some items (e.g. attended EHPS conferences) is higher than in others (e.g. EHP reading frequency).

The items in Figure 2 satisfy the assumptions of the so-called 'congeneric model' of reliability, and the items in Figure 1 satisfy the much more restrictive assumptions of the 'parallel model' of reliability. Just like the differing assumptions of

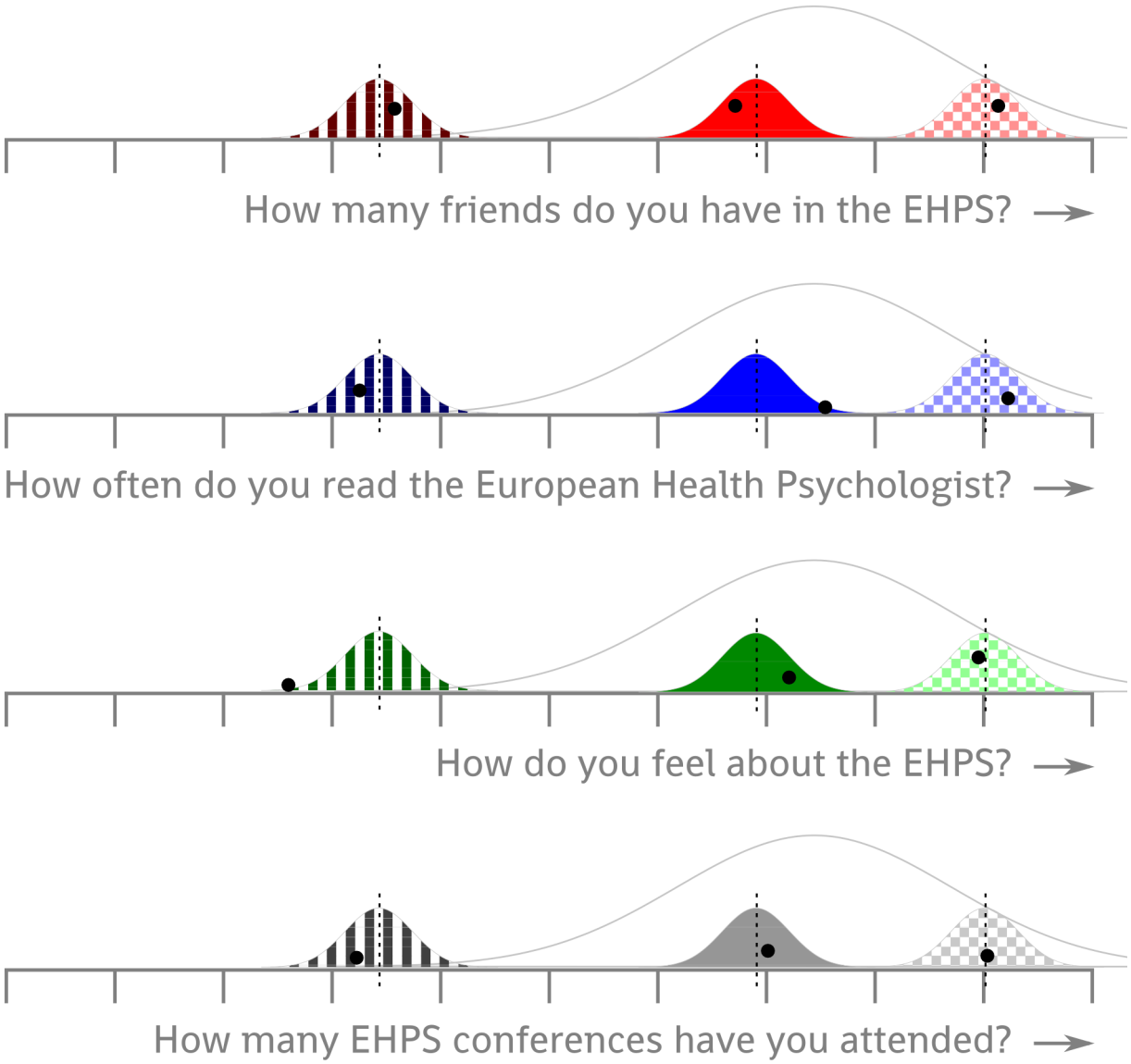


Figure 1: the scores of three individuals on four items that satisfy the assumptions of the 'parallel model of reliability'

the independent samples t-test and the paired samples t-test change the way the value of Student's t needs to be calculated, the assumptions of the different reliability models determine how a test's reliability can be estimated. A shared assumption of both of these models is that the items measure one underlying construct ('unidimensionality'), in this case connectedness with the EHPS. The congeneric model has no additional assumptions, but the parallel model also requires the items to have the same means, the same error variance, and the same variances in and covariances between items. In between this extremely restrictive parallel model and the much more liberal congeneric model lives the 'essentially tau-

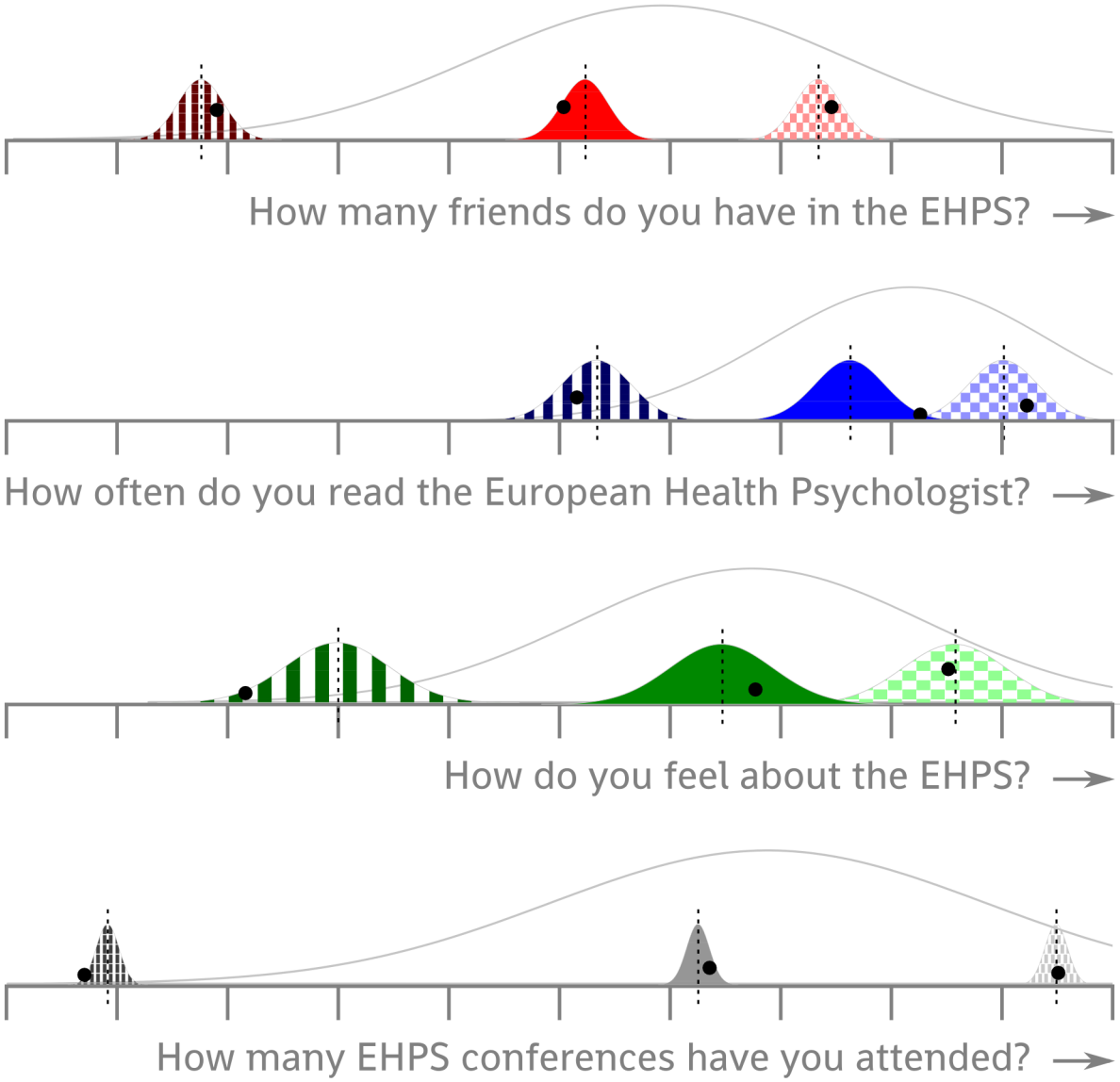


Figure 2: the scores of three individuals on four items that satisfy the assumptions of the 'congeneric model of reliability'

equivalent model', which assumes unidimensionality and equal variances of and covariances between items. This last model is the model relied on by Cronbach's Alpha (Cronbach, 1951). This essentially tau-equivalent model assumes that all items measure the same underlying variable, that they do so on the same scale, and that they are equally strongly associated to that underlying variable. In these situations, Cronbach's Alpha can be calculated as a measure of reliability of the scale; and conversely, violation of these assumptions means that Cronbach's Alpha is no longer a useful measure of reliability. In fact, it can be shown and has been shown that when essential tau-equivalence does not hold, it is

impossible that Cronbach's Alpha equals the reliability of the test (Sijtsma, 2009). Thus, when the assumptions of essential tau-equivalence are violated, the only thing you can be sure of when you know the value of Cronbach's Alpha, is that the test's reliability cannot possibly be that value. Unfortunately, these assumptions are almost always violated in 'real life' (Dunn et al., 2013; Graham, 2006; Revelle & Zinbarg, 2009; Sijtsma, 2009).

Cronbach's alpha is also seen as a measure of a scale's internal consistency, which is often loosely perceived as an indicator of the degree to which the items making up the scale measure the same underlying variable (interestingly, this is the assumption of 'unidimensionality' in the congeneric, parallel, and essentially tau-equivalent models of reliability). However, unfortunately, in addition to the fact that in most situations, Cronbach's Alpha is not a measure of reliability, Cronbach's Alpha has also been shown to be unrelated to a scale's internal consistency. Sijtsma clearly shows that "both very low and very high alpha values can go either with unidimensionality or multidimensionality of the data" (Sijtsma, 2009, p. 119). In other words (almost those of Sijtsma, 2009, p. 107, to be precise): Cronbach's Alpha has very limited usefulness. I therefore recommend that we abandon it.

Aside: note that I have kept these explanations deliberately conceptual. For example, I have conveniently neglected to even acknowledge the semantic swamp that one enters when trying to define reliability and internal consistency (instead, I worked from the assumption that many researchers use Cronbach's Alpha with a vague idea that it provides some information on reliability and/or internal consistency, whatever the precise definitions may be). However, for those readers interested in the technical background to these explanations, an extensive literature is available (Cortina, 1993; Dunn et al., 2013; Graham, 2006; Revelle & Zinbarg, 2009; Sijtsma, 2009). The goal of the current paper is not so much to provide yet another thorough argument of why Cronbach's Alpha should be abandoned; this has been done better than I can by people who understand the issues at hand much better. Instead, this paper is meant to make it easy to adopt a different approach than computing Cronbach's Alpha.

How to abandon Cronbach's Alpha

So, in most situations, we know that if we computed Cronbach's Alpha, the resulting value cannot possibly be the reliability of our scale. This of course begs the question of whether other measures exist that provide better estimates of a scale's reliability. The answer, of course, is yes[1]. Two have been recommended: the 'greatest lower bound' (glb; Sijtsma, 2009) and omega (Revelle & Zinbarg, 2009). Sijtsma (2009) argued that the glb is the lowest possible value that a scale's reliability can have. That means that when the glb is known, the reliability is by definition in the interval [glb, 1]. Revelle and Zinbarg (2009) argue that omega in fact provides a more accurate approximation of a scale's reliability, and that omega is almost always higher. For details on these measures, please see their respective papers; for now, we will focus simply on how to compute these superior estimates of reliability.

Both the glb and omega are available in the free and open source package R (R Development Core Team, 2014), and a step-by-step explanation of how to compute omega has even been published already (Dunn et al., 2013). However, this step-by-step explanation is not

1 Imagine, though, how awkward it would be if I would realise only at this point that none exist...

Open Access, limiting its accessibility to researchers and students. In addition, it involves using quite some R commands, and some researchers have become so accustomed to using SPSS that the idea of learning a new statistical package can seem somewhat daunting. Finally, as we health psychologists know, behavior change is facilitated by making the desired behavior easier to perform. This is where the current paper comes in: it introduces a so-called 'wrapper' function that enables researchers with no knowledge of R to compute a number of measures of reliability with one command. The minor catch is that before this function can be used, it needs to be downloaded and installed into R. However, like downloading and installing R itself, this needs to be done only once; and this, too, consists of only one command. The following paragraphs explain what R is, how to install it, how to install the required package, and how to request the glb and omega.

R is an open source statistical package. It has several advantages over SPSS, such as that it is free and that almost any existing statistical analysis is available. In addition, very accessible introductory texts exist (e.g. Field, Miles, & Field, 2012). It can be downloaded from http://r-project.org. Windows users who prefer to not install anything on their system (or are unable to) can download a portable version from http://sourceforge.net/projects/rportable/, which can even run from a USB stick. Once installed and started, R displays the console, an interface enabling users to input commands for R. The aptly named function 'install.packages' can be used to install packages. Specifically, to install the package we now require, run the following command:

```
install.packages('userfriendlyscience');
```

R will then ask the user to select a mirror. Simply select the geographically closest location, after which R will proceed to download the requested package 'userfriendlyscience' and all packages it depends on. Once the package 'userfriendlyscience' is installed, we need to tell R that we actually require it, using the function 'require', after which we can immediately compute the reliability estimates with 'scaleReliability':

```
require('userfriendlyscience');
scaleReliability();
```

R then presents a dialog where an SPSS datafile can be selected. The function 'scaleReliability' assumes that this datafile only

```
-- STARTING BOOTSTRAPPING TO COMPUTE CONFIDENCE INTERVALS! --
--    (this might take a while, computing 1000 samples)    --
-- FINISHED BOOTSTRAPPING TO COMPUTE CONFIDENCE INTERVALS! --
                 dat: dat.time1
                       Items: all
              Observations: 250
                       Omega: 0.8
Greatest Lower Bound (GLB): 0.85
          Cronbach's alpha: 0.75

Confidence intervals:
                     Omega: [0.74, 0.83]
         Cronbach's alpha: [0.71, 0.79]
```

contains items of one scale. Therefore, before heading into R, store an intermediate version of your datafile from SPSS by selecting the 'Save as …' option in the 'File' menu, in the resulting dialogue clicking the 'Variables…' button, and then using the 'drop' and 'keep' functionalities to select which variables to store. R then produces output similar to that showed at the bottom of the previous page.

Note that after having displayed the first two lines, R starts bootstrapping to generate the confidence intervals, which may take a while. The function scaleReliability has a number of other arguments that can be used, for example to specify which variables in the data should be used, whether to compute confidence interval in the first place, and how many samples to compute for the confidence interval bootstrapping. Interested readers can get more information by entering '?scaleReliability' in the R console. An example script that generates simulated data and computes these estimates (these exact estimates, in fact), as well as the output of the script, is provided at this paper's Open Science Framework page at http://osf.io/tnrxv.

As most researchers know, and as has been argued countless times before, the informational value of point estimates is negligible compared to the value of confidence intervals. However, SPSS does not normally provide confidence intervals for most of the statistics it reports, and this may have contributed to the phenomenon that researchers generally report only a point estimate for their reliability estimates. Hopefully, the fact that scaleReliability by default reports confidence intervals for Omega (and for the old-fashioned researchers among us, for Cronbach's Alpha) can contribute to a change in reporting standards for reliability estimates. Although it would be a huge improvement if researchers would from now on report confidence intervals for omega instead of, or in addition to, point estimates for Cronbach's Alpha, it might be even better to try and decrease our reliance on quantitative 'quality labels' for aggregate measures.

Multidimensional aggregated measures: indices

All measures of reliability discussed here share one important assumption: that of unidimensionality. Even this single assumption, however, is not always plausible. For example, many health psychology studies explore the relative importance of a variety of psychological determinants for predicting a given health behavior. Common determinants included in such studies are attitude, descriptive subjective norm, injunctive subjective norm, and perceived behavioral control. When the study is meant to inform the development of behavior change interventions, these determinants are usually defined as aggregate variables, measured with various items that each reflect a specific belief (Bartholomew, Parcel, Kok, Gottlieb, & Fernández, 2011; Fishbein & Ajzen, 2010). For example, beliefs underlying injunctive subjective norm reflect perceived approval or disapproval of social referents regarding the target behavior; beliefs underlying descriptive norm reflect perceived performance of the target behavior by social referents; and beliefs underlying perceived behavioral control reflect perceived environmental barriers and possessed skills. Imagine, for example, the following three items to measure descriptive norm: "My partner exercises [never-daily]", "My best friend exercises [never-daily]", "Of my colleagues, [none exercise-all exercise]", and the following three items to measure perceived behavioral control: "The sports facility is located [very far-very close] to my home", "For me, exercising

three times a week is [very hard-very easy]" and "A subscription to a sport club is [very expensive-very cheap]".

Most readers will probably feel it coming: these three descriptive norm items do not measure the same dimension, and neither do the perceived behavioral control items. Instead of being meant as repeated measurements of the same underlying unidimensional construct, these items are combined in one measure because aggregating the normative pressure experienced with regards to these different social referents provides a useful indicator of the total pressure experienced. If most of one's colleagues exercise, but one's partner and best friend rarely do, the descriptive norm is considerably lower than when one's partner and best friend also exercise. Similarly, there is no reason to assume that there is a correlation between the proximity of one's house to exercise facilities and one's assessment of the monetary costs of a membership at such facilities; but both measures likely contribute to a person's intention to exercise regularly and their subsequent behavior. Aggregating these measures despite the clear lack of unidimensionality is warranted on the basis of theory: for example, a theory might hold that a person's perceptions of social referents' behavior all influence that person's own intention and behavior in a similar fashion. If a researcher then wants to study the relative contribution of descriptive norms to the prediction of intention and behavior, aggregating these descriptive normative beliefs, which all exert their influence on intention and behavior in a similar manner, makes sense. This allows convenient comparison to the association strength of other determinants such as attitude and perceived behavioral control. To distinguish such deliberately multidimensional aggregate measures from intended unidimensional scales, I will refer to them as indices.

Although for indices, aggregation of the measures can be justified, computation of reliability or internal consistency measures cannot; after all, the assumption of unidimensionality has been violated. Nonetheless, it is not uncommon to see authors computing Cronbach's Alpha for variables such as subjective norm or perceived behavioural control that are measured with items assessing a variety of beliefs. Even worse, in the case of a low value, items might be removed to enhance Cronbach's Alpha, sometimes even causing authors to resort to single-item measures. This means the validity of the relevant measure is decreased on the basis of a flawed measure that should not have been computed in the first place. Of course, for indices, the assumption of the glb or omega would have been violated as well. And to make matters worse more challenging, to a degree this problem of multidimensionality holds for all psychological variables.

Reliability versus validity

The example given above used indices that are commonly adopted in health psychology, and showed how such measures are multidimensional, yet can still be useful aggregate measures. Other psychological variables, such as attitude, coping skills, or optimism, can more easily be argued to be unidimensional. However, even for these constructs, the different items used to measure them are usually not merely intended as exact replications of each other. Besides increasing reliability, a second reason for using multiple measurements to measure a construct is increased validity. Take for example these three items from the General Self-Efficacy (GSE) scale, all answered on a 4-point scale from "Not at all true" to "Exactly true": "I can always manage to

solve difficult problems if I try hard enough", "If someone opposes me, I can find the means and ways to get what I want" and "It is easy for me to stick to my aims and accomplish my goals" (see e.g. Luszczynska, Scholz, & Schwarzer, 2005). Each of these items taps quite different aspects of self-efficacy: the first item concerns self-efficacy regarding difficult problems, and imposes the condition of considerable investment of resources; the second item concerns general self-efficacy, but only under the circumstances where another person attempts to thwart goal-directed behavior; and the third item taps both self-efficacy and perceived self-regulatory skill. These three aspects are different, but all are part of the generic construct general self-efficacy. The GSE scale contains these items not to enhance reliability, but to enhance validity of the scale.

The fact that measures such as the GSE contain items that measure different aspects of a construct is not a weakness of the measure: rather, it is a strength. Very narrowly defined and measured psychological constructs have very limited applicability; in fact, most psychological constructs derive part of their usefulness from the generic level at which they are defined. For example, the Reasoned Action Approach recommends applying the principle of compatibility when measuring behavior and its determinants (Fishbein & Ajzen, 2010). This principle assumes that any behavior has four defining elements (action, target, context, and time), and dictates that behavior and its determinants must be measured with regards to the exact same action, target, context, and time. For example, when measuring EHPS conference attendance and its determinants, an intention item might be "Will you attend the EHPS conference in 2014? [absolutely not-absolutely]", a subjective norm item might be "How many of your colleagues will attend the EHPS conference in 2014? [none-all]", and a self-efficacy item might be "How easy or hard will it be for you to attend the EHPS conference in 2014? [very easy-very hard]". The measure of self-efficacy acquired this way will have extremely high applicability when predicting EHPS conference attendance in 2014, but it will be almost useless for anything else (such as predicting exercise behavior). By contrast, general self-efficacy is useful to predict a broad range of behaviors precisely because of its generic nature. Thus, many psychological constructs derive their usefulness from their relatively broad definition, and therefore, their relatively broad operationalization.

At the same time, the fact that different aspects of a psychological construct are measured means that the measure can never be perfectly unidimensional. Although an individual's response to each item should normally be determined mainly by the psychological construct of interest, other psychological constructs will have an influence as well; and accordingly, factor analysis may reveal that the first factor explains a disappointingly low proportion of variance. However, this does not have to be a problem: after all, if a set of items measures a very generic psychological construct, influence of related psychological constructs is to be expected. Scale diagnostics cannot be interpreted without taking into account how specific or generic the measured construct is defined. Therefore, scale inspection should entail more than computation and evaluation of a single quantitative measure.

A comprehensive assessment of scale quality

If we acknowledge that aggregate measures contain different items to enhance both

reliability and validity, and that more specific, more narrowly operationalized measures are not by definition better than more generic, more broadly operationalized measures, it becomes even harder to defend thresholds for estimates such as Cronbach's Alpha. Even when refraining from relying on tentative thresholds, Cronbach's Alpha, omega, and the glb provide only a very narrow view on the dynamics of a scale. In addition, it seems useful to examine the degree of unidimensionality of a scale by conducting a factor analysis (or principal component analysis, depending on the goal), and inspecting the Eigen values of each component, as well as the factor loadings. Furthermore, inspecting the distribution of each item, as well as the way the items are associated, can help identify anomalies in single measures. Therefore, I suggest that researchers routinely generate a combination of diagnostics:

1. Compute omega, the glb, and Cronbach's alpha, preferably with confidence intervals;
2. Conduct a factor analysis or principal component analysis and inspect all Eigen values and the factor loadings (at least for the first factor);
3. Inspect the means, medians, and variances for each item;
4. Generate a correlation matrix;
5. Inspect the scatterplots of the associations between all items;
6. Inspect histograms of each item's distribution.

These diagnostics should then be interpreted in conjunction with the separate measurements of the aggregate measure (e.g. the complete list of the items forming a scale in a questionnaire). Unfortunately, inspecting such a diverse combination of diagnostic information means that providing clear guidelines as to when a scale is acceptable becomes impossible. Of course, that was more or less the point of this contribution: because operationalization and measurement are so important to psychological science, assessment of successful operationalization deserves more attention than simple comparison to a quantitative threshold. Conveniently, the R package described above just so happens to contain another function called 'scaleDiagnosis', which provides most of these diagnostics. It can be used the same way 'scaleReliability' is used:

```
scaleDiagnosis();
```

The user can then select an SPSS datafile, after which the function produces output similar to that shown on the next page. The function also creates a plot similar to the one shown in Figure 3[2]. This so-called scattermatrix shows the (bivariate) scatterplots of the combinations of all items in the scale, as well as the univariate distribution of each item, and the point estimates for the correlation coefficients in the upper right half. This is useful for quick visual inspection of the nature of the associations between the items and their distributions. This output, the text as text file and the plot both as .png and .svg, is also available at this paper's Open Science Framework page at http://osf.io/tnrxv.

However, forgoing the comfort of a quantitative threshold means that decisions about scale construction become much more subjective. It seems wrong to on the one hand acknowledge the importance and complexity of these decisions, and on the other hand, forgo the convenient possibility of external scrutiny that quantitative measures such as Cronbach's Alpha seem to afford. And indeed, this would be wrong. The problems of the so-called 'researcher degrees of freedom' have been made painfully clear recently (Simmons, Nelson, & Simonsohn, 2011), and the solution is straightforward: as argued before in the European Health Psychologist, researchers should fully disclose

```
                dat: res$dat
                    Items: t0_item1, t0_item2, t0_item3, t0_item4, t0_item5
              Observations: 250
                    Omega: 0.8
Greatest Lower Bound (GLB): 0.85
          Cronbach's alpha: 0.75

Eigen values: 2.924, 0.64, 0.566, 0.463, 0.407
Loadings:
         PC1
t0_item1 0.76
t0_item2 0.78
t0_item3 0.75
t0_item4 0.78
t0_item5 0.75

                PC1
SS loadings    2.92
Proportion Var 0.58

         var   n  mean   sd median trimmed   mad   min   max range  skew kurtosis   se
t0_item1   1 250 17.94 2.92  18.00   17.97  2.92 10.93 24.68 13.75 -0.08    -0.54 0.18
t0_item2   2 250 31.70 9.52  31.05   31.66  9.06  2.62 58.49 55.87  0.04    -0.01 0.60
t0_item3   3 250 20.26 3.53  19.99   20.26  3.71 11.81 30.20 18.39  0.06    -0.51 0.22
t0_item4   4 250 34.48 5.99  34.14   34.32  5.80 20.14 56.36 36.22  0.32     0.21 0.38
t0_item5   5 250 29.78 9.73  29.63   29.69 10.09  4.60 56.98 52.38  0.08    -0.16 0.62
```

(Peters, Abraham, & Crutzen, 2012). In this case, such disclosure would mean making these diagnostics public, along with the complete questionnaires that were used. Preferably, these resources are published in an Open Access online repository such as the free Open Science Framework (see https://osf.io/), as this makes them available to the entire scientific community. This will be of considerable use to other researchers who are constructing similar measurement instruments. At the very least, researchers should publish these scale diagnostics as supplementary materials with their articles[3]. Publishing the scale diagnostics will enable reviewers to critically and thoroughly assess the integrity of the used measurement instruments, and can facilitate both interpretation of the findings and future meta-analysis.

The question then becomes, what do we know about the quality of measurement instruments of studies that only report Cronbach's Alpha? The answer is, very little. We know that the reliability is in any case not the value reported for Cronbach's alpha (but by definition something higher, although we have

2 To store a plot in R, the 'Save as' option in the 'File' menu can be used.

3. Although this is less desirable, as it will restrict access to this information if the main article is behind a paywall.

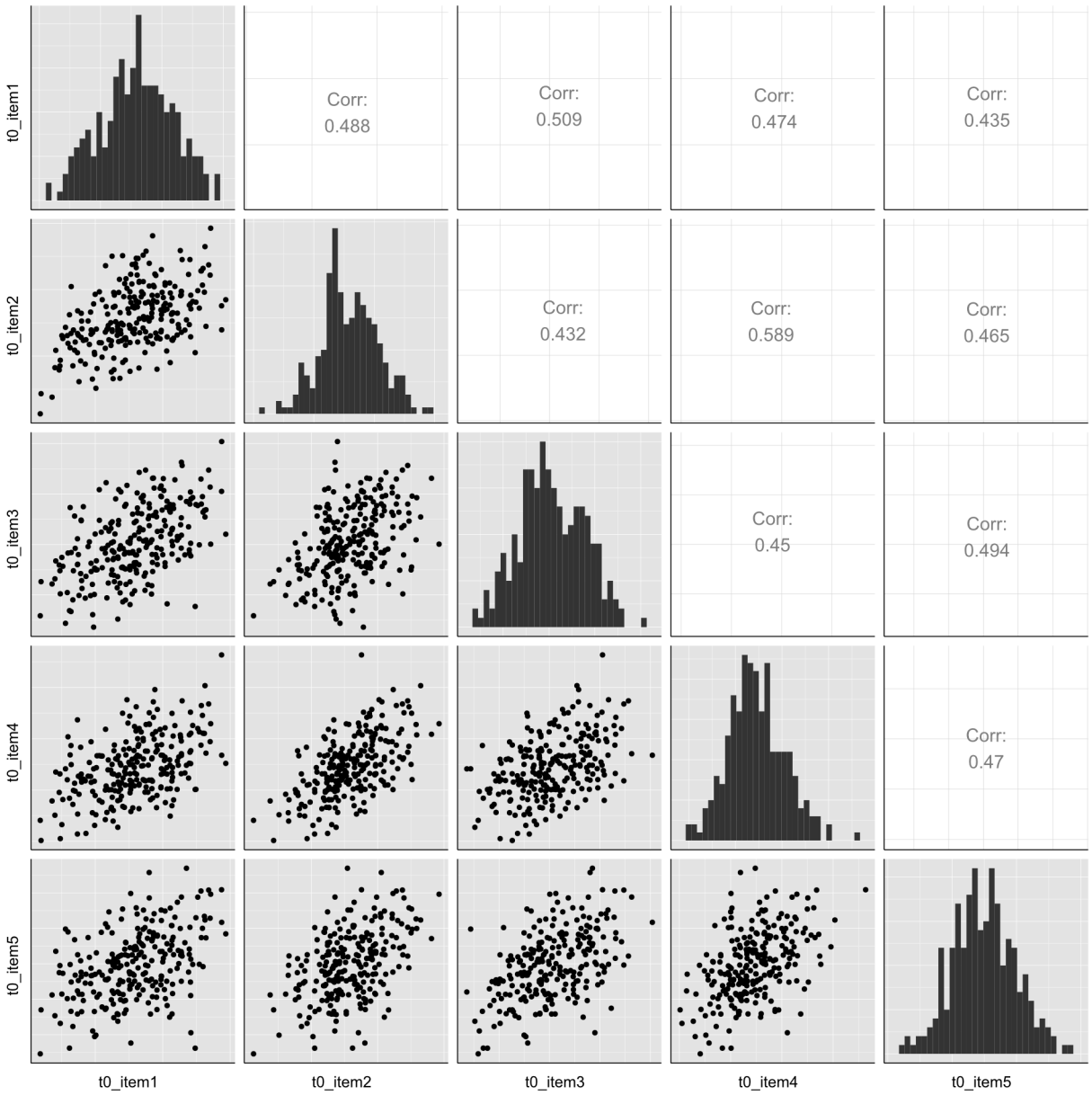


Figure 3: A scattermatrix as produced by the scaleDiagnosis() function in the userfriendlyscience package for R

no clue as to how much higher). We know nothing about the internal consistency of the scale. For those studies that published the questionnaires as appendices or supplemental materials, it is possible to inspect the items to establish the face validity (i.e. whether the items seem to tap cognitions/emotions that make up or contribute to the construct the scale intends to measure); and if correlation tables were published as well, a more thorough assessment of the measurement instruments becomes possible. However, without such information, we know almost nothing about the validity and reliability of the used measures. If we assume that the validity and reliability of the measurement instruments used in most

studies are acceptable, the only remaining problem is that we don't know which studies are the ones with unacceptable measures.

Conclusion

Researchers often compute and report Cronbach's Alpha to determine whether aggregate measures have acceptable reliability or internal consistency. Although most authors and reviewers seem content with this, Cronbach's Alpha is both unrelated to a scale's internal consistency and a fatally flawed estimate of its reliability. In addition, this reliance on one quantitative estimate fails to acknowledge the relationship between reliability and validity. Finally, some measures are deliberately multidimensional (indices), violating the assumption of unidimensionality underlying Cronbach's Alpha, omega and the Greatest Lower Bound. Scale diagnostics would be improved if researchers would assess, simultaneously, estimates and their confidence intervals for omega, the glb, and perhaps Cronbach's Alpha; Eigen values and factor loadings; individual item distributions; and a correlation- and scattermatrix of all items. These diagnostics should be assessed in conjunction with the raw measurement instrument (e.g. the items in a scale). This will enable researchers to base their decisions on a more complete picture of scale performance. In addition, publishing these diagnostics and the measurement instruments will enable reviewers and readers to closely scrutinize the reliability and validity of such measures. Finally, such a process will enable considerable acceleration of scale construction in general, as it will become possible to spot and study item formulations that consistently perform badly. It is important not to underestimate the importance of how we measure our psychological variables of interest, since psychologists do not have the luxury of the more objective measures that many other disciplines use (after all, even implicit and biopsychological measures are indirect and require many assumptions). Hopefully, this paper and the R functions described herein will have made it sufficiently easy for this more comprehensive assessment of scale quality to become commonplace.

References

Bartholomew, L. K., Parcel, G. S., Kok, G., Gottlieb, N. H., & Fernández, M. E. (2011). *Planning health promotion programs: an Intervention Mapping approach* (3rd ed.). San Fran-cisco, CA: Jossey-Bass.

Cortina, J. M. (1993). What is Coefficient Alpha? An examination of theory and application. *Journal of Applied Psychology, 78(1), 98–104. doi:10.1037/0021-9010.78.1.98*

Cronbach, L. J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. doi:10.1007/BF02310555

Dunn, T. J., Baguley, T., & Brunsden, V. (in press). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*. doi:10.1111/bjop.12046

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics using R*. London: Sage Publications Ltd.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: the reasoned action approach*. New York: Psychology Press.

Graham, J. M. (2006). Congeneric and (Essentially) Tau-Equivalent Estimates of Score Relia-bility: What they are and how to use them. *Educational and Psychological Measurement, 66*(6), 930–944. doi:10.1177/0013164406288165

Luszczynska, A., Scholz, U., & Schwarzer, R. (2005). The general self-efficacy scale:

multicul-tural validation studies. *The Journal of Psychology, 139*(5), 439–57. doi:10.3200/JRLP.139.5.439-457

Peters, G.-J. Y., Abraham, C. S., & Crutzen, R. (2012). Full disclosure: doing behavioural science necessitates sharing. *The European Health Psychologist, 14*(4), 77–84.

R Development Core Team. (2014). *R: A language and environment for Statistical Computing*. Vienna, Austria. Retrieved from http://www.r-project.org/

Revelle, W., & Zinbarg, R. E. (2009). Coefficients Alpha, Beta, Omega, and the glb: comments on Sijtsma. *Psychometrika, 74*(1), 145–154. doi:10.1007/s11336-008-9102-z

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika, 74*(1), 107–120. doi:10.1007/s11336-008-9101-0

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–66. doi:10.1177/0956797611417632

Gjalt-Jorn Y. Peters
is Assistant professor of Psychology at the Open University of the Netherlands, Heerlen, the Netherlands

gjalt-jorn@behaviorchange.eu

commentary

Time is a jailer: what do alpha and its alternatives tell us about reliability?

Rik Crutzen
Maastricht University

Psychologists do not have it easy, but the article by Peters (2014) paves the way for more comprehensive assessment of scale quality. I plead guilty to habitually reporting alpha[1] and – in some cases where I did not – reviewers were so kind to request this as well. Peters (2014) rightly states that alpha is a fatally flawed estimate of a scale's reliability. He presents readily available alternatives, such as the greatest lower bound (glb) or omega, as superior estimates of reliability. I agree with the suggestion by Peters (2014) to routinely generate a combination of diagnostics, but I think we are still missing out on an important aspect of reliability: test-retest reliability.

Figure 1 might bring flashbacks to your Statistics 101-course. My apologies if this side effect is an unpleasant experience. The figure is very useful, however, to explain test-rest reliability. Whereas Peters (2014) discusses items within a scale (e.g., attitude items), I will focus on the scale in its entirety (e.g., an attitude measure). So, each dot in Figure 1 represents, for example, a single administration of an attitude measure. The closer these dots are to the bull's eye, the more likely that the scale actually measures attitude. This concerns the validity of the scale. However, if we use the same measure repeatedly over time, we also want to be sure that we get the same score (if nothing has changed). So, the dots should be close to each other (or, ideally, overlap each other). This is an aspect of a scale's reliability.

A legitimate question to ask is why time is such an important factor contributing to reliability? The reason behind this is that over time both true scores and measurement error can fluctuate. The observed test score (e.g., a participant's score on an attitude measure) is the sum of the true score (e.g., a participant's actual attitude) and the measurement error. This measurement error does not only differ between participants or items within a scale, but also within participants over time (Guttman, 1945). At the same time, however, differences in the observed test can also be the result of actual changes in attitude.

Imagine an intervention targeted at the attitude towards use of protective clothing to prevent tick bites (see e.g., Crutzen & Beaujean, 2014 for brief background information). The efficacy of this intervention is tested in a two-arm randomized controlled trial (RCT) with a waiting-list control group. No change is expected in the control group, and differences in the intervention group should reflect differences in true scores regarding attitude. This does not mean that test–retest reliability is only desirable in measures of constructs that are not expected to change over time. It can be, for example, that a national health campaign about prevention of tick bites is launched during the trial period. This might lead to changes in attitude of the control group as well. Therefore, an important aspect of assessing scale quality is

1 Cronbach considered it an embarrassment that the formula became conventionally known as Cronbach's alpha (Cronbach & Shavelson, 2004).

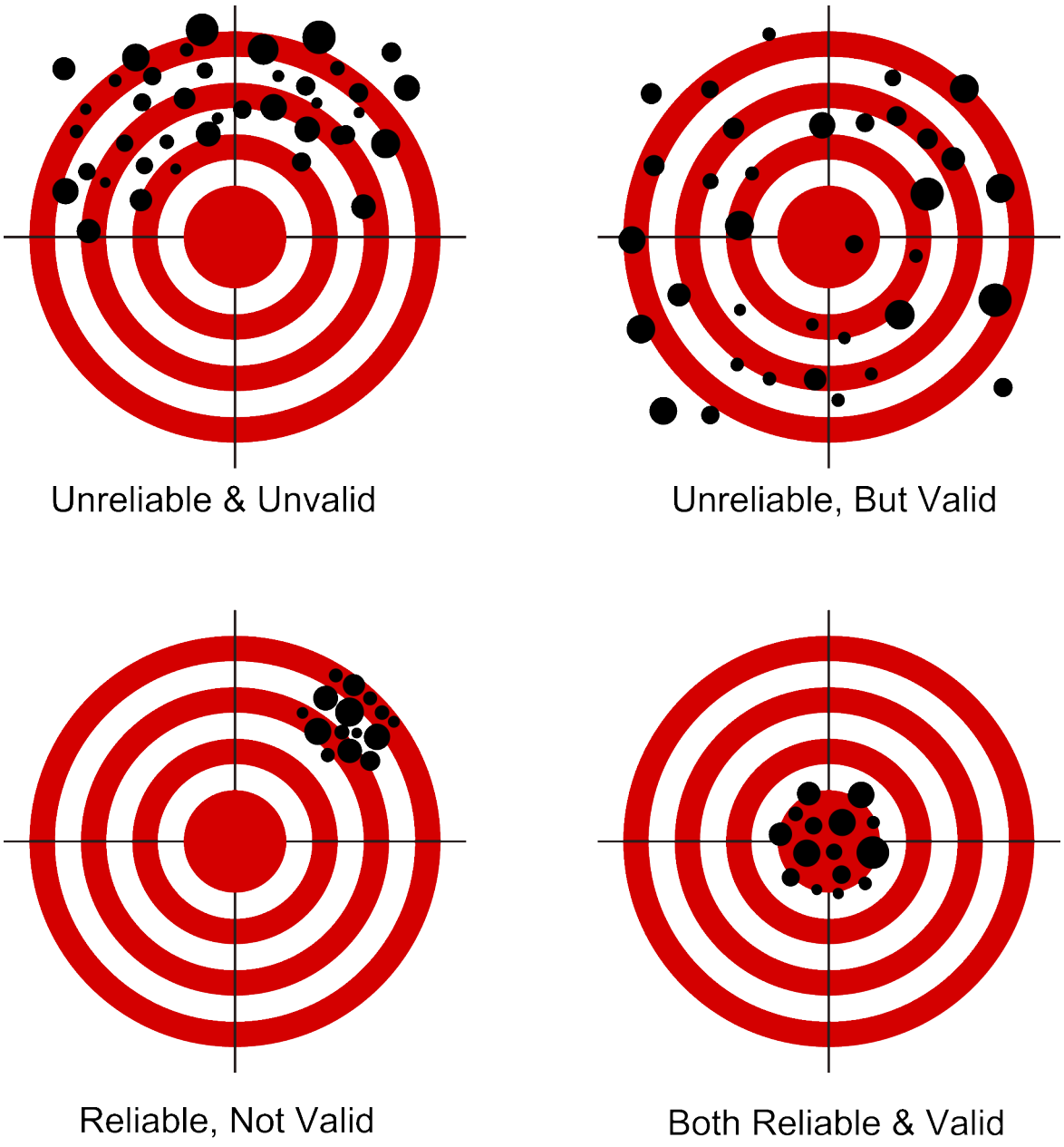


Figure 1. Reliability and validity (© Nevit Dilmen).

to distinguish between changes in observed scores due to actual changes in, for example, attitude (even if they are unexpected) and changes in measurement error due to time (also known as transient error).

The magnitude of transient error in real data can range from non-existent to very large (Becker, 2000). Ignoring transient error can lead to inaccurate conclusions (Chmielewski & Watson, 2009). Even though I agree with the suggestions by Peters (2014), they are not sufficient to address transient error. It appears that high internal consistency does not indicate that a scale can measure change reliably, nor can it estimate stability of true scores. The opposite is also true; a low internal consistency does not attenuate stability (McCrae, Kurtz, Yamagata, & Terracciano, 2011).

Sijtsma (2009) questions the estimation of reliability on the basis of a single administration of a scale, even when using alternatives to alpha such as the glb. Nevertheless, there are indices that go beyond traditional correlate coefficients and that explicitly take transient error into account. Green (2003), for example, explains an index based on coefficient alpha, but susceptible to transient error, whereas Schmidt, Le, and Ilies (2003) present a procedure for estimating the coefficient of equivalence and stability (CES). Test-retest data is required for these indices. Huysamen (2006) argues that "the very reason for the original coefficient's popularity has been that it doesn't require a retest, and Green's coefficient has to forgo this luxury, as any other index that wishes to reflect transient error by definition has to do."

This leaves us at a crossroad. We more or less ignore transient error and simply go on[2] or we agree that test-retest analyses should be part of comprehensive assessment of scale quality. In case of the latter, we have to acknowledge that this brings additional workload. This additional workload does not only concern the need for test-retest data, but assessing test-retest reliability also brings along additional issues. For example, the choice of an appropriate retest interval[3] (Chmielewski & Watson, 2009; Green, 2003) and comparison of the indices between domains (Schmidt et al., 2003). I hope that the arguments presented in this article are convincing to take on this additional workload.

To make this workload as minimal as possible, I will now briefly explain how to easily compute these indices. First, install R and the package 'userfriendlyscience' (see Peters, 2014). Then, in your commonly used statistical environment (e.g., SPSS), create a data file that only contains the items of the two administrations of your scale. The order is important: the items of the first administration should come first, followed by, in the same order, the items of the second administration (e.g., "t0_item1", "t0_item2", and "t0_item3" followed by "t1_item1", "t1_item2", and "t1_item3"). Then, load this data file into R and compute the test-retest alpha coefficient and the CES with:

```
testRetestReliability();
```

R again opens a dialog to enable selection of the data file, after which output similar to the below will be shown.

Note that computing the single administration indices (e.g., original coefficient alpha, omega, and the glb, computed in Peters, 2014) yielded much higher values (.75-.85). This means that when using this scale and computing single-administration reliability indices, one might erroneously assume a negligible effect of transient error, which might have far-reaching consequences in non-experimental designs.

In the ideal situation, we choose to conduct test-rest analyses as part of comprehensive assessment of scale quality, but how do we achieve this? A (too) simple, but nonetheless recommendable, first step would be to conduct test-retest analyses whenever longitudinal data are available (e.g., after conducting an RCT). It would be far better to conduct a pre-test to assess test-retest reliability, using the indices mentioned above. In such a pre-test, the choice of retest interval should be grounded theoretically depending on the construct of interest (Chmielewski & Watson, 2009). This

2 Following the suggestions by Peters (2014) is already a big step forward.
3 E.g., a personality trait measure might be less likely to change over time in comparison with an attitude measure.

```
                    Items at time 1: t0_item1, t0_item2, t0_item3, t0_item4, t0_item5
                    Items at time 2: t1_item1, t1_item2, t1_item3, t1_item4, t1_item5
                        Observations: 250
          Test-retest Alpha Coefficient: 0.43
Coefficient of Equivalence and Stability: 0.45

To help assess whether the subscales (automatically generated using means) are parallel, here
are the means and variances:
Mean subscale a1, time 1: 82.19 (variance = 235.07)
Mean subscale a2, time 1: 51.95 (variance = 132.18)
Mean subscale a1, time 2: 83.57 (variance = 243.19)
Mean subscale a2, time 2: 52.09 (variance = 138.76)
```

might all sound as "yet another thing to do before I can run my study". However, if we agree on the importance of pre-testing our intervention materials to avoid counterproductive results (e.g., Whittingham et al., 2009), I think we should be as strict with regard to the measures of constructs we are interested in. After all, we draw our conclusions based on these measures and we should not try "to explain findings that result from transient error masquerading as true change" (Chmielewski & Watson, 2009).

Acknowledgements

I would like to thank Gjalt-Jorn Y. Peters for adding the test-retest alpha coefficient and the CES to the userfriendlyscience R package.

References

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods, 5*(3), 370-379. doi:10.1037/1082-989X.5.3.370

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: the impact of transient error on trait research. *Journal of Personality and Social Psychology, 97*(1), 186-202. doi:10.1037/a0015618

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391-418. doi:10.1177/0013164404266386

Crutzen, R., & Beaujean, D. (2014). Preventive behaviours regarding tick bites. *BMJ, 348*, g231. doi:http://dx.doi.org/10.1136/bmj.g231

Green, S. B. (2003). A coefficient alpha for test-retest data. *Psychological Methods, 8*(1), 88–101. doi:10.1037/1082-989X.8.1.88

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255-282. doi:10.1007/BF02288892

Huysamen, G. K. (2006). Coefficient alpha: unnecessarily ambiguous; unduly ubiquitous. *SA Journal of Industrial Psychology, 32*(4), 34-40. doi:10.4102/sajip.v32i4.242

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15*(1), 28-50. doi: 10.1177/1088868310366253

Peters, G.-J. Y. (2014). The alpha and the omega of scale reliability and validity: why and how

to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *The European Health Psychologist, 16*(2), 54-67

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: an empirical examination of the effects of different sources of measurement error on realibility estimates for measures of individual differences constructs. *Psychological Methods, 8*(2), 206-224. doi:10.1037/1082-989X.8.2.206

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107-120. doi:10.1007/s11336-008-9101-0

Whittingham, J. R. D., Ruiter, R. A. C., Bolier, L., Lemmers, L., Van Hasselt, N., & Kok, G. (2009). Avoiding counterproductive results: an experimental pretest of a harm reduction intervention on attitude toward party drugs among users and nonusers. *Substance Use & Misuse, 44*(4), 532-547. doi:10.1080/10826080802347685

Rik Crutzen
is Assistant Professor at the Department of Health Promotion, Maastricht University, Maastricht, The Netherlands

rik.crutzen@maastrichtuniversity.nl

original article

Bayesian analyses: where to start and what to report

Rens van de Schoot
Utrecht University & North-West University
Sarah Depaoli
University of California

Most researchers in the social and behavioral sciences will probably have heard of Bayesian statistics in which probability is defined differently compared to classical statistics (probability as the long-run frequency versus probability as the subjective experience of uncertainty). At the same time, many may be unsure of whether they should or would like to use Bayesian methods to answer their research questions (note: all types of conventional questions can also be addressed with Bayesian statistics). As an attempt to track how popular the methods are, we searched all papers published in 2013 in the field of Psychology (source: Scopus), and we identified 79 empirical papers that used Bayesian methods (see e.g. Dalley, Pollet, & Vidal, 2013; Fife, Weaver, Cool, & Stump, 2013; Ng, Ntoumanis, Thøgersen-Ntoumani, Stott, & Hindle, 2013). Although this is less than 0.5% of the total number of papers published in this particular field, the fact that ten years ago this number was only 42 indicates that Bayesian methods are slowly beginning to creep into the social and behavioral sciences.

The current paper aims to get you started working with Bayesian statistics. We provide: (1) a brief introduction to Bayesian statistics, (2) arguments as to why one might use Bayesian statistics, (3) a reading guide used to start learning more about Bayesian analyses, and, finally (4) guidelines on how to report Bayesian results. For definitions of key words used in this paper, please refer to Table 1.

Bayesian Statistics: A brief introduction

Before providing arguments why one would use Bayesian statistics, we first provide a brief introduction. Within conventional statistical techniques, the null hypothesis is always set up to assume no relation between the variables of interest. This null hypothesis makes sense when you have absolutely no idea of the relationship between the variables. However, it is often the case that researchers do have *a priori* knowledge about likely relationships between variables, which may be based on earlier research. With Bayesian methods, we use this background knowledge (encompassed in what is called a 'prior') to aid in the estimation of the model. Within Bayesian statistics, we can learn from our data and incorporate new knowledge into future investigations. We do not rely on the notion of repeating an event (or experiment) infinitely as in the conventional (i.e., frequentist) framework. Instead, we incorporate prior knowledge and personal judgment into the process to aid in the estimation of parameters.

Thus, the key difference between Bayesian statistics and conventional (e.g., maximum likelihood) statistics concerns the nature of the unknown parameters in a statistical model. The unknown model parameters are those that are freely estimated. For example, when estimating a regression model with one dependent outcome variable (Y) and two predictors (X1 and X2), see Figure 1, the unknown parameters are: one

Table 1: *A brief definition of key words and phrases*

Key Words and Phrases	*Definition*
Background Knowledge	Knowledge about population parameter values (e.g., a regression coefficient) that can be determined based on prior research, an analysis of previous data, or expert opinions.
Bayesian Statistics	A statistical tool that can be used to combine background knowledge of population parameters with current data to obtain estimates via the resulting posterior distribution.
Credibility Interval	The Bayesian version of the traditional confidence interval. Can be interpreted as the (e.g. 95%) probability that the population parameter is between the particular upper and lower bounds determined by the Bayesian credibility interval.
Confidence Interval	Frequentist (conventional) confidence intervals are based on repeated sampling theory such that, for a 95% confidence interval, 95 out of 100 replications of exactly the same experiment capture the fixed but unknown regression coefficient.
Frequentist Statistics	A class of statistics that relies on point estimation and opposes Bayesian statistics because it does not incorporate background knowledge into the estimation process (e.g., maximum likelihood estimation methods).
Hyperparameters	The specific parameters for a prior distribution. For example, if a normal distribution is selected for the prior, then the mean and variance parameters of this normal prior are called the hyperparameters. The values specified for the hyperparameters control the amount of (un)certainty incorporated into the model about a given parameter.
Likelihood Function	Represents the observed data likelihood. This weights the prior distribution in Bayesian statistics to obtain the posterior distribution form which we draw inferences.
Markov chain Monte Carlo (MCMC)	A simulation-based estimation method that is used to make simulated draws from a distribution and form a Markov chain that represents the posterior distribution.
p-value	In frequentist statistics, this is the probability of obtaining a test statistic as or more extreme than the critical value, given that the null hypothesis is true.
Parameter	A fixed but unknown feature of the model that is estimated either through frequentist or Bayesian methods.
Posterior	The distribution that is obtained once combining the prior and the likelihood in the Bayesian estimation process.
Posterior p-value	Bayesian p-value that is based on the posterior distribution obtained.
Precision	The amount of information incorporated into a prior distribution. More information is equated to having a larger degree of precision (less uncertainty) and therefore equates to smaller variability in the prior. Precision is specifically defined as the inverse of the variance.
Prior	A statistical distribution that can be used to capture the amount of (un)certainty in a population parameter. This distribution is then weighted by the sample data to obtain the posterior, which is used to make inference.

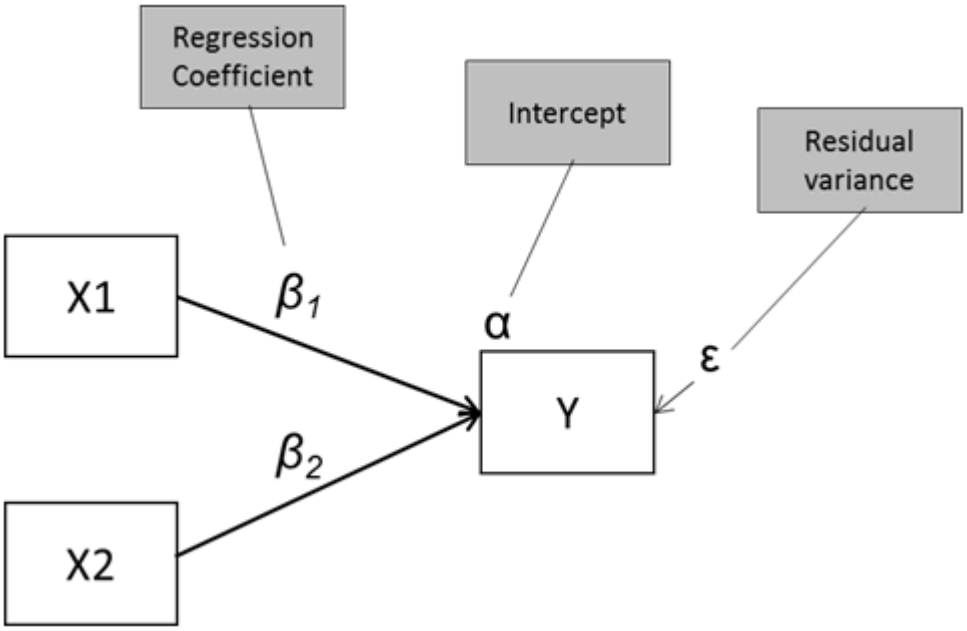


Figure 1. Regression model with the unknown parameters.

intercept (α), two regression coefficients (β_1, β_2), and the residual variance of the dependent variable (ε). With conventional statistics it is assumed that in the population there is only one true population parameter, for example, one true regression coefficient that is fixed but unknown. In the Bayesian view of probability, all unknown parameters can incorporate (un)certainty that can be defined by a probability distribution. Thus, Bayesian methods do not provide one outcome value but rather an interval ('distribution') with a probability that this interval contains the regression coefficient. That is, each parameter is believed to have a distribution that captures (un)certainty about that parameter value. This (un)certainty is captured by a distribution that is defined *before* observing the data and is called the *prior distribution* (or *prior*). Next, the *observed* evidence is expressed in terms of the *likelihood function* of the data. The data likelihood is then used to weigh the prior and this product yields the *posterior distribution*, which is a compromise of the prior distribution and the likelihood function. These three ingredients constitute the famous Bayes' theorem.

The three ingredients underlying Bayesian statistics are summarized in Figure 2 for one of the regression coefficients (β) pulled from Figure 1. The first ingredient of Bayesian statistics is knowledge about this parameter before observing the data, as is captured in the prior distribution. Often this knowledge stems from systematic reviews, meta-analyses or previous studies on similar data (see O'Hagan et al., 2006). In Figure 2 five different priors are displayed for β. The variance, or precision (inverse of the variance), of the prior distribution reflects one's level of (un)certainty about the value of the parameter of interest: the smaller the prior variance, the more certain one is about the parameter value. There are three main classes of priors that differ in the amount of certainty they carry about the population parameter. These different priors are called: (1) non-informative priors, (2) informative priors, and (3) weakly-informative priors. Non-informative priors are used to reflect a great

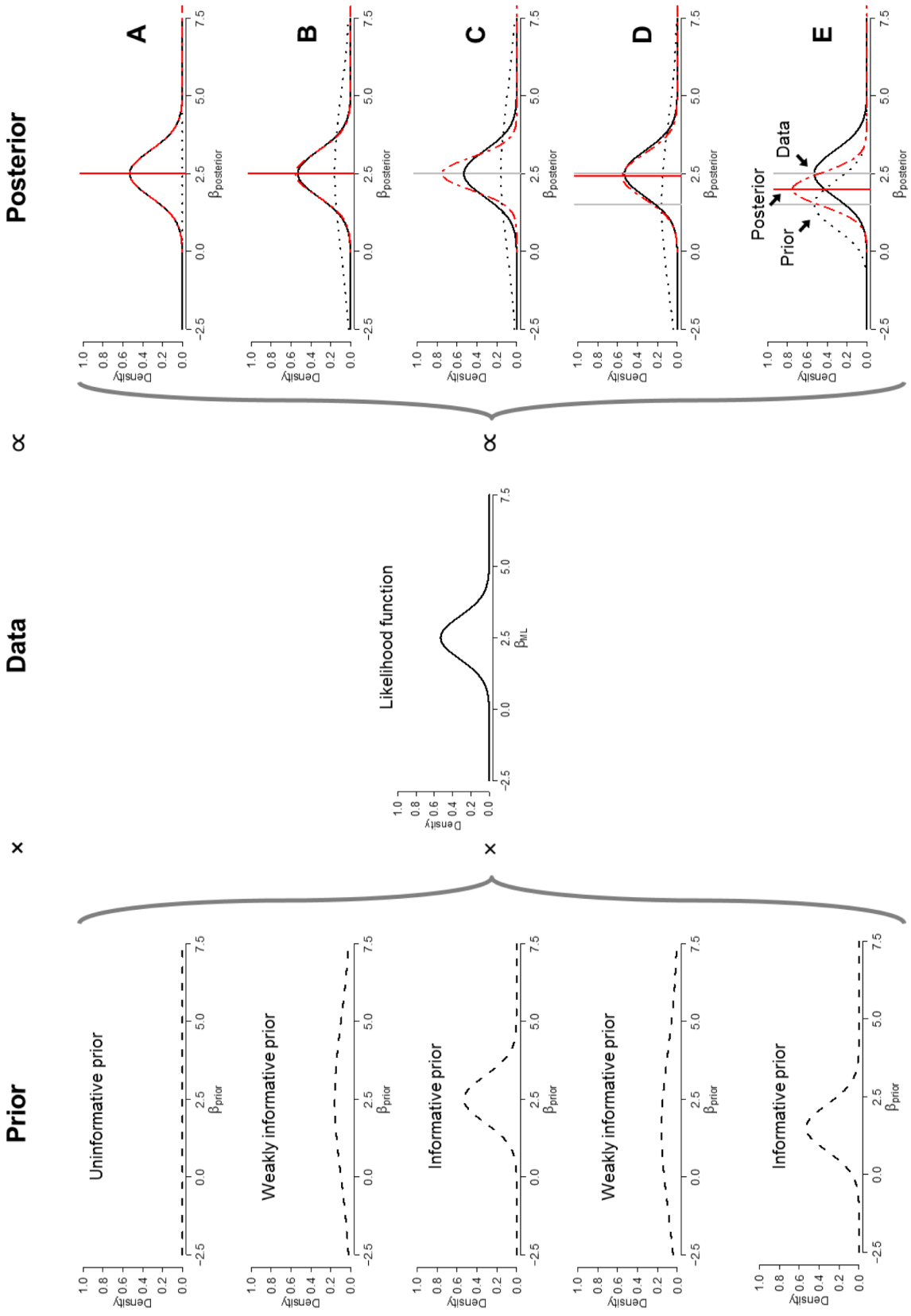


Figure 2. The three ingredients of Bayesian statistics for the regression coefficient.

deal of uncertainty in what the population parameter looks like. Weakly-informative priors incorporate some information into the model and reflect more certainty about the population parameter compared to a non-informative prior. This prior contains some useful information, but it does not typically have much influence on the final parameter estimate. Finally, the prior that contains the most amount of certainty about the population parameter is an informative prior. Informative priors contain strict numerical information that is crucial to the estimation of the model and can have a large impact on final estimates. These three levels of informativeness are created by modifying the parameters of the prior, called hyperparameters. Specifically, the hyperparameters for these priors (e.g., the prior mean and prior variance) are fixed to express specific information and levels of (un)certainty about the model parameters being estimated.

The second ingredient is the information in the data itself. It is the observed evidence expressed in terms of the likelihood function of the data (β_{ML}). Thirdly, both *prior* and *data* are combined via Bayes' theorem. The *posterior distribution* reflects one's updated knowledge, balancing background knowledge (the prior) with observed data (the likelihood). With a non or weakly informative prior, the posterior estimate may not be influenced by the choice of the prior much at all, see Figure 2A, 2B and 2C. With informative (or subjective) priors, the posterior results will have a smaller variance, see Figure 2C. If the prior disagrees with the information in the data, the posterior will be a compromise between the two, see Figure 2E, and then one has truly learned something new about the data or the theory.

Why would one use Bayesian Statistics?

There are four main reasons as to why one might choose to use Bayesian statistics: (1) complex models can sometimes not be estimated using conventional methods, (2) one might prefer the definition of probability, (3) background knowledge can be incorporated into the analyses, and (4) the method does not depend on large samples.

First, some complex models simply cannot be estimated using conventional statistics. In these cases of rather complex models, numerical integration is often required to compute estimates based on maximum likelihood estimation, and this method is intractable due to the high dimensional integration needed to estimate the maximum likelihood. For example, conventional estimation is not available for many multilevel latent variable models, including those with random effect factor loadings, random slopes when observed variables are categorical, and three-level latent variable models that have categorical variables. As a result, alternative estimation tools are needed. Bayesian estimation can also handle some commonly encountered problems in orthodox statistics. For example, obtaining impossible parameters estimates, aiding in model identification (Kim, Suh, Kim, Albanese, & Langer, 2013), producing more accurate parameter estimates (Depaoli, 2013), and aiding in situations where only small sample sizes are available (Zhang, Hamagami, Wang, Grimm, & Nesselroade, 2007).

Second, many scholars prefer Bayesian statistics because of the different definition of probability. Consider for example the interpretation of confidence intervals (CIs). The frequentist CI is based on the assumption of a very large number of repeated samples from the

population. For any given data set, a regression coefficient can be computed. The correct frequentist interpretation for a 95% CI is that 95 out of 100 replications of exactly the same experiment capture the fixed but unknown regression coefficient. Often this 95% CI is misinterpreted as meaning there is a 95% probability that the regression coefficient resides between the upper and lower limit, which is actually the Bayesian interpretation. Thus, Bayesian confidence intervals may be more intuitively appealing.

Third, as described above, with Bayesian statistics one can incorporate (un)certainty about a parameter and update this knowledge. Let background knowledge be the current state of affairs about a specific theoretical model, which can be operationalized by means of a statistical model, see for example Figure 1. Everything that is already known about the parameters in the model based on, for example, previous publications, can be used to specify informative priors, see Figure 2. When the priors are updated with current data, something can be learned, especially if the priors (i.e., current state of affairs) disagree with the current data. Let us explain this conflict between the prior and the current data using a simplified example where two groups were generated (M_1=0, M_2=0.45, SD=2; n=100) using an exact data set. Obviously, when no prior knowledge is specified (using non-informative prior distributions), there is no difference between the population difference ($\Delta M_{population}$ = 0.45) and the estimated difference obtained with the Bayesian analysis ($\Delta M_{posterior}$ = 0.45). Next, we specified informative priors that were inaccurate to the population; that is, for M_1 we specified a prior mean of .50 and for M_2 we specified a prior mean of .05. We varied the precision of the prior distribution to obtain weakly informative (low precision) and highly informative priors (high precision). The relation between the precision and the prior-data conflict (i.e., the difference between $\Delta M_{population}$ and $\Delta M_{posterior}$) is shown in Figure 3. In conclusion, the higher the precision, the more influence the prior specification has on the posterior results. If there is a large prior-data conflict, apparently the current state of affairs about the statistical model does not match with the current data. This is what a Bayesian would call: fun! Because

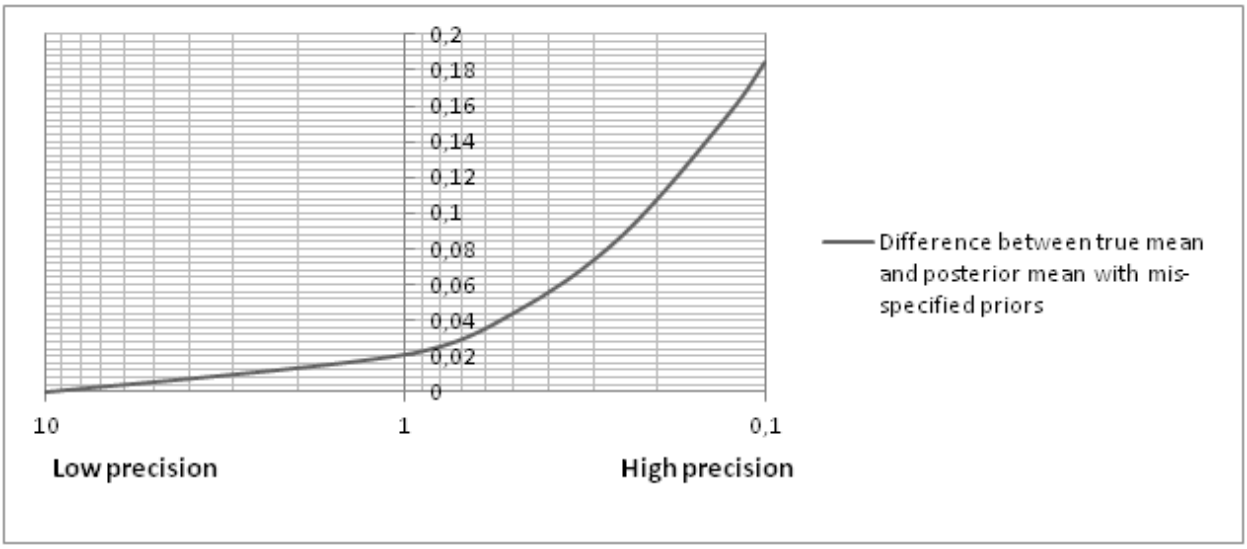


Figure 3. The relation between precision and bias.

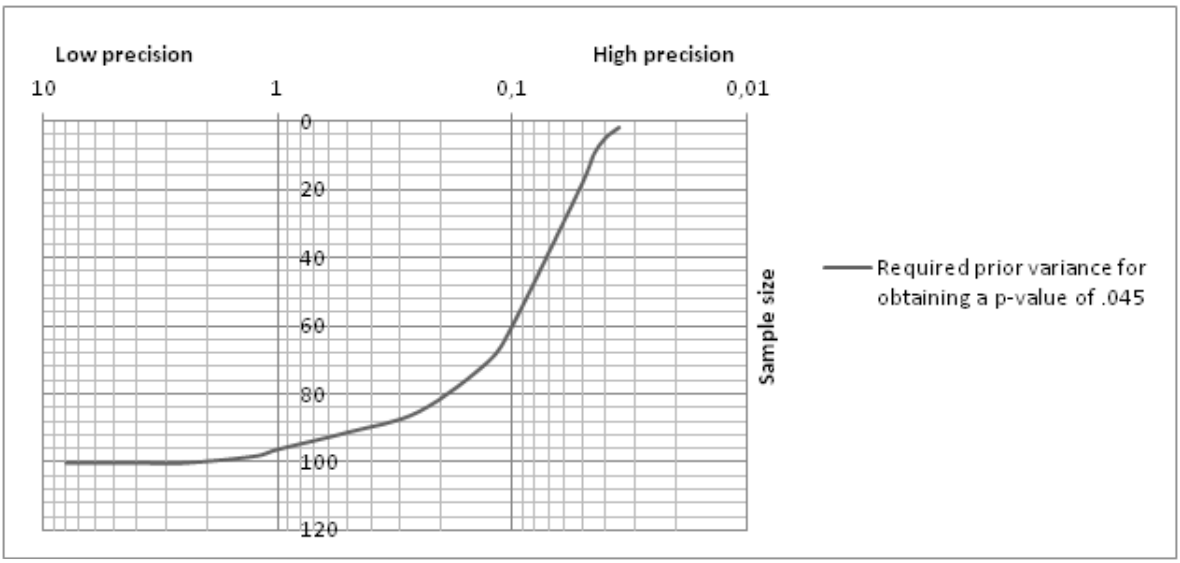


Figure 4. The relation between precision and the possible gain in sample size.

now, finally, something new has been discovered and one should discuss in the paper how it could be that there is a prior-data conflict. Is it the theory that needs to be adjusted? Or, was the data not a random sample from the population? Or does the theory not hold for the specific population used for the current study? All of these questions are related to updating knowledge.

Fourth, Bayesian statistics is not based on large samples (i.e., the central limit theorem) and hence large samples are not required to make the math work. Many papers have shown the benefits of Bayesian statistics in the context of small data set (e.g., Zhang et al., 2007).To illustrate the decrease in required sample size we performed a small simulation study. Multiple exact data sets with two groups, see above, were generated with the goal to obtain for every data set the same *p*-value for a *t*-test. With n = 100 the *t*-test produced a just significant effect of p = .045. Also, when using objective Bayesian statistics with an infinitive low prior precision (non-informative prior) the Bayesian *p*-value was .045. Next, we specified weakly and highly informative priors with a prior mean equal to the population values (data based prior), but we varied the precision. The relation between the precision and the required sample size to obtain the same significant effect of p = .045 is shown in Figure 4 showing that the higher the precision, the smaller the sample size needed to obtain the same effect. In conclusion, the more precision a researcher is willing to specify before seeing the data, the smaller the sample size needed to obtain the same effect compared to an analysis without specifying any prior knowledge.

1 When using exact data sets the data characteristics are exactly the same as the population statistics. For example, if the population mean is specified as being zero with a standard deviation of 2, the data set generated from this population also has exactly a mean of zero and a SD of 2. The software BIEMS (Mulder, Hoijtink, & de Leeuw, 2012) was used for generating such an exact data. The t-tests for mean differences were performed in the software Mplus.

Where to start?

Of course, the introduction offered in the current paper is not enough to start working with Bayesian statistics, therefore we provide a step-by-step reading guide as well as resources for statistical programs that can implement

Bayesian methods. For a gentle introduction to Bayesian estimation, we recommend the following: Kaplan and Depaoli (2013); Kruschke (2011); and van de Schoot et al. (2013). For a more advanced treatment of the topic, readers can be referred to a variety of sources, which include Gelman, Carlin, Stern, and Rubin (2004).

There are many different software programs that can be used to implement Bayesian method in a variety of contexts and we list the major programs here. Various packages in the R programming environment (e.g., Albert, 2009) implement Bayesian estimation, with the number of Bayesian packages steadily increasing. Likewise, AMOS (Arbuckle, 2006), BUGS (Ntzoufras, 2009), and M*plus* (Muthén, 2010) can be used for estimating Bayesian latent variable models, which can also include multilevel or mixture extensions. BIEMS (Bayesian inequality and equality constrained model selection; Mulder, Hoijtink, & de Leeuw, 2012) is a Bayesian program for multivariate statistics and Bayesian hypothesis testing. Standard statistical models estimated through the SAS software program can now be used for Bayesian methods. Finally, SPSS incorporates Bayesian methods for imputing missing data.

What to include in an empirical Bayesian paper?

There are several key components that must be included in the write-up of an empirical paper implementing Bayesian estimation methods. The statistical program used for analysis is an important detail to include since different methods (called *sampling methods*) are implemented in different Bayesian programs and these methods may lead to slightly different results. A discussion of the priors needs to be in place. The researcher should thoroughly detail and justify all prior distributions that were implemented in the model, even if default priors were used from a software program. It is important to always provide these details so that results can be replicated, a full understanding of the impact of the prior can be obtained, and future researchers can draw from (and potentially update) the priors implemented. A discussion of chain convergence must be included. Each model parameter estimated should be monitored to ensure that convergence was established for the posterior. A variety of statistical tools can be used to help monitor and evaluate chain convergence (see, Sinharay, 2004), and visual inspection of convergence plots can also aid in detecting non-convergence. Finally, researchers might also find it beneficial to run a sensitivity analysis using different forms and levels of informativeness for the priors implemented. Although we do not recommend using this as a means for updating the prior on the same data set (i.e., the original prior should still be used in the final write-up), the sensitivity analysis can help provide insight into the impact of the prior and this impact can be discussed further in the paper.

Conclusion

In our experience, we have found Bayesian methods to be incredibly useful for solving estimation problems, handling smaller sample sizes with greater accuracy, and incorporating prior judgment or knowledge into the estimation process. It is our aim that this paper will serve as a starting point for those interested in implementing Bayesian methods.

Author's note

The first author was supported by a grant from the Netherlands organization for scientific

research: NWO-VENI-451-11-008.

References

Albert, J. (2009). *Bayesian computation with R*. New York: Springer.

Arbuckle, J. L. (2006). *Amos* (Version 7.0) [Computer Program]. Chicago: SPSS.

Dalley, S. E., Pollet, T. V., & Vidal, J. (2013). Body size and body esteem in women: The mediating role of possible self expectancy. *Body Image, 10*(3), 411-414. doi:10.1016/j.bodyim.2013.03.002

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods, 18*(2), 186-219. doi:10.1037/a0031609

Fife, B. L., Weaver, M. T., Cook, W. L., & Stump, T. T. (2013). Partner interdependence and coping with life-threatening illness: The impact on dyadic adjustment. *Journal of Family Psychology, 27*(5), 702-711. doi:10.1037/a0033871

Gelman, A.,Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman & Hall.

Kaplan, D. & Depaoli, S. (2013). Bayesian statistical methods. In T. D. Little (ed.), *Oxford handbook of quantitative methods* (pp 407- 437). Oxford: Oxford University Press.

Kim, S. Y., Suh, Y., Kim, J. S., Albanese, M., & Langer M. M. (2013). Single and multiple ability estimation in the SEM framework: a non-informative Bayesian estimation approach. *Multivariate and Behavioral Research, 48*(4), 563-591. doi:10.1080/00273171.2013.802647

Muthén, B. (2010). *Bayesian analysis in Mplus: A brief introduction*. Technical Report. Version 3.

Mulder, J., Hoijtink, H., & de Leeuw, C. (2012). BIEMS: A Fortran 90 program for calculating Bayes factors for inequality and equality constrained models. *Journal of Statistical Software, 46*(2), 1-39.

Ng, J. Y. Y., Ntoumanis, N., Thøgersen-Ntoumani, C., Stott, K., & Hindle, L. (2013). Predicting psychological needs and well-being of individuals engaging in weight management: The role of important others. *Applied Psychology: Health and Well-being, 5*(3), 291-310. doi:10.1111/aphw.12011

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, NJ: Wiley.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J.,... Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. West Sussex: Wiley.

Sinharay, S. (2004). Experiences with Markov Chain Monte Carlo Convergence Assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics, 29*(4), 461-488. doi:10.3102/10769986029004461

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J. & van Aken, M. A. G. (2013). A gentle introduction to Bayesian Analysis: Applications to research in child development. *Child Development*. doi:10.1111/cdev.12169

Zhang, Z., Hamagami, F., Wang, L., Grimm, K. J., & Nesselroade, J. R. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development, 31*(4), 374-383. doi:10.1177/0165025407077764

Rens van de Schoot
is Assistant Professor at Utrecht University, the Netherlands and extra-ordinary Professor at the Optentia research programme, North West University, South-Africa
a.g.j.vandeschoot@uu.nl

Sarah Depaoli
is Assistant Professor at University of California, Merced, USA
sdepaoli@ucmerced.edu