Intensive Measurement Designs for Research on Aging

Rast, P., MacDonald, S.W.S., Hofer, S.M.

2012

https://doi.org/10.1024/1662-9647/a000054

Published in final edited form as:
GeroPsych (Bern). 2012 ; 25(2): 45–55. doi:10.1024/1662-9647/a000054.

Intensive Measurement Designs for Research on Aging

Philippe Rast, **Stuart W. S. MacDonald**, and **Scott M. Hofer**
Department of Psychology, University of Victoria

Abstract

Intensive measurement burst designs permit analysis of behavioral and biological processes as they unfold over short and long periods of time and providing the opportunity to identify change from an individual's normative level of functioning. The measurement burst design permits statistical decomposition of short-term variation and learning effects that overlay normative aging and provide stronger bases for detecting accelerated change due to pathological processes. We provide an overview of design features and analysis of measurement burst data in Project MIND. The objective of intensive measurement designs is to obtain greater resolution of processes of interest that permit reliable and sensitive assessments of functioning and change in functioning and of key determinants underlying short-term variation and long-term aging and health-related change.

Keywords

multilevel modeling; longitudinal methods; individual differences; variability; developmental change

Innovation in study design and within-person measurement is essential for progress in detecting cognitive aging and identifying the determinants of such changes, such as stress and lifestyle factors and health-related processes such as dementia and cardiovascular disease. A variety of research designs have been used within the broad enterprise of lifespan developmental and aging research, ranging from cross-sectional and longitudinal quasiexperimental designs to randomized control trials. Observational (i.e., quasiexperimental) studies are the most typical ones because many questions and hypotheses focus on constructs that cannot be manipulated, due to practical and ethical reasons.

However, several long-standing and major questions of interest to the gerontological community have not been and cannot be answered with existing data based on cross-sectional or widely spaced longitudinal study designs. For example, the question concerning when aging-related cognitive decline begins has been addressed using both age-heterogeneous cross-sectional and longitudinal designs, yielding different conclusions depending on the type of data that were analyzed. That is, results obtained from cross-sectional data have typically identified negative age-related differences in the mid-twenties (e.g., Salthouse, 2009), while results from longitudinal data generally point to age-related declines after the age of 55 (e.g., Schaie, 1996).

Philippe Rast: Department of Psychology, University of Victoria, PO BOX 3050, Victoria, BC V8W 3P5, Canada, Tel. +1 250 472–4869, prast@uvic.ca.

Declaration of Conflicts of Interest
The authors declare that no conflicts of interest exist.

Several features of these different study designs can explain the discrepancy in the findings. Cross-sectional results, for example, are confounded by birth cohort effects (Rönnlund & Nilsson, 2008; Schaie, 2008) and mortality selection (e.g., Kurland, Johnson, Egleston, & Diehr, 2009), which greatly limit the opportunities to draw inferences about individual and population change (Hofer, Flaherty, & Hoffman, 2006; Hofer & Sliwinski, 2001; Kraemer, Yesavage, Taylor, & Kupfer, 2000; Molenaar, Huizenga, & Nesselroade, 2003; Wohlwill, 1973). A major example of this is the Flynn effect, a consistent finding of secular increases in cognitive capabilities over the last 70 years in many different countries (Flynn, 1984, 1987; Rönnlund & Nilsson, 2008).

Given these limitations, it is now widely accepted that understanding aging requires observing and evaluating longitudinal data that enables one to estimate changes within individuals (see Bauer, 2011). However, longitudinal designs are susceptible to testing effects due to the repeated presentation of stimuli and previous test exposure of the participants, leading to downwardly biased estimates of change rates and later onset of cognitive decline (Ferrer, Salthouse, Stewart, & Schwartz, 2004; Salthouse, 2011; Thorvaldsson, Hofer, Berg, & Johansson, 2006).

Despite the potential for bias related to retest and reactivity, longitudinal designs remain the most adequate way to capture and explain things relative within person changes and development. In combination with modeling techniques such as multilevel modeling (Bryk & Raudenbush, 1992), this type of design provides a number of insights into developmental and aging-related processes which were not achievable with cross-sectional designs. For example, we are able to address questions about differences in cognitive decline and about predictors of these changes on an individual level. However, the typical design of longitudinal studies is one based on widely spaced assessments that typically span several years between the measurement occasions. Such studies lack temporal resolution on the lower scale and are not sensitive to within-person dynamics, changes, or events that occur between assessments. However, a number of longitudinal studies of aging have implemented daily diary or other forms of intensive measurements to augment the widely spaced assessments. These alternative longitudinal designs provide measurement occasions on different temporal scales (e.g., across days and years) as well as a basis for disentangling short-term repeated testing effects from long-term within-person change (Hofer, Rast, & Piccinin, 2011; Hoffman, Hofer, & Sliwinski, 2011; Sliwinski, Hoffman, & Hofer, 2010).

A number of questions cannot be definitively answered with typical widely spaced longitudinal studies. For example, a question of importance might be whether changes in cognition occur at the same time or consecutively and if there are distinct patterns of cognitive change and variation which indicate different causal processes? Does it all go together when it goes (Rabbitt, 1993) or is cognitive decline a diversified and highly individual process different for every cognitive domain and every individual? This also means that change over time should not only be observed in the average performance but also in the covariances and variances among a set of variables (e.g., Zimprich & Rast, 2009). Further, one might be interested in whether psychological phenomena, which are observed on one timescale, can inform us about phenomena on a different timescale. For example, one question might be whether variability in a reaction time task predicts long-term changes in cognition (see MacDonald, Hultsch, & Dixon, 2003).

More generally, the question might be whether there are processes that can be observed at a short timescale which reproduce on a longer timescale and can therefore serve as an early prediction of later changes? Specifically, adaptation over a short timeframe such as *learning* might share components with adaptation over a longer timeframe such as *development.* One might hypothesize that individual differences in the ability to learn might also account for

individual differences in cognitive aging, although less in an explanatory sense, but rather as a parallelism of development taking place in a different timeframe. Several researchers have speculated that learning might represent "microdevelopment," that is, development within a short timeframe, as opposed to "macrodevelopment," which typically covers development over longer timespans (Hultsch, 1974; Yan & Fischer, 2002). Similarly, Lindenberger and Baltes (1995) speculated that the mechanisms underlying learning might be similar to those underlying cognitive development, thus turning the study of learning into a showcase examining cognitive development (cf., Fischer, 1980; Hultsch, 1974). Although development and learning are often treated as a dichotomy, both are characterized by a persistent change of behavior over time, albeit across different timescales (Newell, Liu, & Mayer-Kress, 2001). Moreover, one might argue that both share important principles, because variability, selection, and adaptation are central to change within individuals. In addition, questions regarding within-person variability cannot typically be addressed in a multiwave longitudinal setting (Ram & Gerstorf, 2009), and call for special designs such as the intensive measurement design which is the focus of the present work.

Intensive Measurement Designs

A relatively recent and underused approach that might prove to be fruitful for answering or approaching some of these questions is the measurement burst design (Nesselroade, 1991b; Nesselroade & McCollam, 2000; Sliwinski, 2008; Walls, Barta, Stawski, Collyer, & Hofer, 2011), which features sets of measurements comprised of a number of closely spaced assessments. That is, one burst might consist of multiple assessments that take place on a daily or weekly basis. The bursts themselves are spaced over longer intervals such as, for example, months or weeks. This design allows one to effectively separate short-term (e.g., day-to-day) within-person variability from long-term (year-to-year) within-person level, change, and variation. This type of design can be seen as a special case of the interrupted time-series design (see Walls et al., 2011), with the distinction that the interruptions are planned and based on theoretical and empirical decisions regarding the within-person variation and change in the outcomes of interest. The strength of the burst design is the ability to improve the precision of the estimate by measuring a variable repeatedly over a short period of time. By doing so, the power for the detection of long-term change in burst studies can be increased, and they can be designed with shorter intervals and fewer subjects compared to multiwave designs.

Long-term widely spaced longitudinal designs confound retest effects and long-term change. Measurement burst designs, in turn, provide a more complete separation of different temporal processes and have been used to separate short-term retest effects from long-term age-related cognitive change (e.g., Sliwinski et al., 2010). They have also been used to evaluate how within-person emotional reactivity to stress (across a week of repeated measures) is related to decreases in cognitive performance (e.g., Sliwinski, Smyth, Hofer, & Stawski, 2006). Hofer, Rast, and Piccinin (2012), for example, show how a repeated series of measurements can be decomposed into a population average trend, individual deviation from the trend, and time-specific variation about the individual-level slope as distinct from error. Note that in most designs the error is indistinguishable from reliable time-specific variation (but see Rast, Hofer, Sparks, 2012).

This is also a specific problem of multiwave studies that fail to take into account inherent intraindividual variability (Nesselroade, 1991a), in the sense that the measurements within the waves are assumed to adequately capture the location of the individual. If this is not the case, inference from longitudinal designs may be seriously biased. As Sliwinski (2008) showed, this problem is related to the reliability of the outcome variables. Personality states, for example, might be reliably measured but exhibit high systematic within-person variation

and therefore show low retest reliability. Low reliability and high within-person variation can lead to biased estimates of long-term change when the number of occasions is relatively few (e.g., three occasions). The strength of the burst design is the ability to improve the precision of the occasion-specific score by aggregating the within-person distribution of scores across a short period of time. By doing so, the power for the detection of long-term change in burst studies can be increased, and they can be designed with shorter intervals and fewer subjects compared to multiwave designs.

Choice of Temporal Sampling

A critical point in burst designs is the choice of temporal sampling, both within and across bursts. Different time sampling schedules will yield patterns of variability resulting from different influences on the individual (e.g., Martin & Hofer, 2004): sampling over seconds or minutes (e.g., attentional lapses); within test (e.g., practice gains); within session (e.g., fatigue, order effects, motivation); within day (e.g., diurnal effects); across days or weeks (e.g., stress, context, daily variation in health, practice); or across months or years (e.g, development, aging, chronic health conditions).

Sensitive Detection of Within-Person Change and Adaptive Longitudinal Designs

The sampling interval should be chosen according to the research question with more frequent measurements when change is expected to occur and less frequent measurement when stability is expected. This approach might work as long as we are, for example, observing normal cognitive change where we expect accelerated decline between the ages of 60 and 80 years. However, change is a within-person process that occurs at different rates and at different points in time for different people. As such, within-person change may not be well captured by static, widely spaced multiwave designs - even more so for nonnormative change, such as disease-related cognitive decline, where it is hardly possible to define a priori the most adequate sampling interval for each individual.

Consider, for example, the estimation of change points in cognitive performance, an outcome of particular interest for research on dementia (e.g., Hall, Lipton, Sliwinski, & Stewart, 2000; Hall et al., 2007; Jacqmin-Gadda, Commenges, & Dartigues, 2006): The idea is that the identification of within-person change-points (i.e., deviation from an established within-person normative trajectory), in the context of an individual's characteristic level and magnitude of short-term variation, might function as an early marker of disease-related processes. The main challenge in these studies involves the precise estimation of the change point that might happen to be within two widely spaced measurement occasions. The quality of the estimation of the change point is linked to the amount of data that was available around that given point in time and to the variability of the data points, higher variability being related to less precise change point estimates. In an optimal design, one would sample more frequently around the change point in order to enhance the precision of that estimate. Given that the change point itself is unknown, this possibility remains hypothetical and can only be examined with sufficient follow-up before and after the change-point, indicating acceleration in change.

Extant evidence suggests that reaction time variability is related (Hultsch, MacDonald, & Dixon, 2002; MacDonald et al., 2003) or even precedes (Lövdén, Li, Shing, & Lindenberger, 2007) cognitive decline in older adults. This information might be well extracted from bursts and used to define the frequency of prospective, additional measurements taken in between the fixed measurement bursts in an individually centered design. That is, the burst design might be expanded to an adaptive longitudinal burst design where participants are tested on an "as-needed" basis. If the aim is to observe changes as they occur, we need to be able to add additional observations to enhance the precision of the

parameter estimates. This also means that a shift to more adaptive testing procedures is necessary. For example, a burst design might be used to establish a baseline for each participant: Less consistent (i.e., more variable) participants need to be sampled more often than participants with very consistent responses on a given variable. Once a baseline is set, consecutive measures follow in order to monitor changes in the consistency of the responses given by each individual. Considering the findings that variability in reaction times indicates changes in cognition, the sampling interval could be reduced if the variability increases as it may indicate nonnormative changes in cognition.

This speculative example illustrates how empirically derived markers of change may be used to define adaptively the optimal design to capture accelerations or decelerations in cognitive change within individuals. Certainly, issues of repeated testing and practice effects need to be considered in this regard, although such factors may be mitigated in an intensive measurement design where change in maximal performance is the outcome of interest.

Change in Asymptote and Structure of Learning Effects

A major outcome of intensive measurement, particularly measurement burst designs, is the opportunity to model long-term change and change points based on an “asymptotic” estimate of an individual’s functioning (e.g., Hoffman et al., 2011; Sliwinski et al., 2010) that is distinguished from short-term gains (i.e., learning, retest effects) and variation. These types of analyses are limited to data that are collected in measurement burst designs. In the following section, we provide an empirical example of this analysis and the utility of information gained from longitudinal designs sampling from different temporal intervals.

Example Using Data from Project MIND

Sample Description and Procedure

One of the greatest challenges in intensive measurement designs concerns how such data are analyzed. Obviously, different questions require different analysis methods; here, we present one possibility for the examination of data from a burst design. Specifically, measurement sessions within each burst are nested, so that the data points are not independent and require special statistical analyses. Multilevel models are frequently employed in order to take data dependency into account (e.g., Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002). Given that intensive measurement design data are collected repeatedly, we typically have three or more levels of analysis: Several sessions (Level 1) that are nested in consecutive bursts (Level 2), which are again nested within persons (Level 3).

To illustrate how data from an intensive measurement burst design can be evaluated, we use data from the Project MIND (e.g., Hultsch, Strauss, Hunter, & MacDonald, 2008), a study based on 304 community-dwelling adults (208 women and 96 men) ranging in age from 64 to 92 years (M = 74.02, SD = 5.95). They were recruited through advertisements in the local media (newspaper and radio) requesting healthy community-dwelling volunteers who were concerned about their cognitive functioning. All participants were Caucasian. Exclusionary criteria included a diagnosis of dementia by a physician or a Mini Mental Status Examination (MMSE; Folstein, Folstein, & McHugh, 1975) less than 24, a history of significant head injury (defined as loss of consciousness for more than 5 minutes), other neurological or major medical illnesses (e.g., Parkinson’s disease, heart disease, cancer), severe sensory impairment (e.g., difficulty reading newspaper-size print, difficulty hearing a normal conversation), drug or alcohol abuse, a current psychiatric diagnoses, psychotropic drug use, and lack of fluency in English.

For the present work, we selected the first four bursts, each 1 year apart. In the first year, the burst contained five sessions, and in the following 3 years the bursts contained four sessions. In each burst, the sessions within the bursts are scaled in terms of weeks, with the bursts scaled in years.

Materials

Digit Symbol

Processing speed was assessed using the WAIS-R Digit Symbol Substitution task (Wechsler, 1981). Participants were presented with a coding key pairing nine numbers (1 through 9) with nine different symbols. Printed under the coding key were rows of randomly ordered numbers with empty boxes below. Participants were given 90 s to transcribe as many symbols as possible into the empty boxes based on the digit-symbol associations specified in the coding key. The number of correctly completed items represented the outcome measure.

Choice Reaction Time 1-Back (CRT)

This RT task served as the dependent variable in our analyses. Participants received a warning stimulus consisting of a horizontal row of four plus signs on the screen. The response keyboard had four keys in a horizontal array corresponding to the display on the screen. After a delay of 1 s, one of the plus signs changed into a box. The location of the box was randomly equalized across trials. Participants were instructed to press the key corresponding to the location of the box on the previous trial as quickly as possible. Although the instructions emphasized speed, participants were also instructed to minimize errors. A total of 10 practice trials and 61 test trials were administered. Because participants made no response on Trial 1, the latencies and percent correct of the remaining 60 test trials were actually used for analysis.

Statistical Analyses

In order to make use of the measurement burst design, which expands over a longer time-frame, we used a multilevel approach where the different timescales are incorporated simultaneously. That is, in the Project MIND data we have up to four levels where the first-level addresses repeated measurement (trials) in each CRT on a scale of seconds. Given that we are more interested in larger scales such as weeks and years, we will not specifically address the serial position of the 60 reaction time trials within a task. Hence, in this case level 1 captures the outcome across each session within a burst, level 2 captures crossburst burst variation, and level 3 describes the individual as the hierarchically highest level. Given that level 3 captures the individual, it describes the between-person (BP) variability across all levels. The variability in levels 1 and 2 describe the within-person (WP) variability across weeks and years, respectively. Formally, following the notation introduced by Raudenbush and Bryk (2002), the three levels may be defined as

$$\begin{aligned}
&\text{Level 1:} && Y_{ijk}=\pi_{0jk}+\pi_{1jk}Session_{ijk}+\pi_{2jk}Session_{ijk}Burst_{jk}+\varepsilon_{ijk}\\
&\text{Level 2:} && \pi_{0jk}=\beta_{00k}+\beta_{01k}Burst_{jk}+r_{0jk}\\
& && \pi_{1jk}=\beta_{10k}+r_{1jk}\\
&\text{Level 3:} && \beta_{00k}=\gamma_{000}+u_{00k}\\
& && \beta_{01k}=\gamma_{010}+u_{01k}\\
& && \beta_{10k}=\gamma_{100}+u_{10k}
\end{aligned} \tag{1}$$

where Y_{ijk} is the observed value for person *k*, in burst *j*, and session *i*. At Level 1 the slope π_{1jk} is defined by the session scaled in weeks. Further, we need to add a crosslevel

interaction term π_{2jk} to account for the possibility that changes (slope) in RT's across the sessions change on average for each burst. That is, the gain in RT due to weekly exposure to the test might be strongest in the first burst and smallest in the last burst. Hence, it makes a difference whether we are looking at the slope in the first burst compared to, for example, the slope in the fourth burst. The residuals are captured in the error term ε_{ijk} with

$$\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2) \quad (2)$$

At level 2, the slope β_{01k} is defined by the burst and its time metric is years. Here, we define the first set of random effects *r* which operate at the within-burst level. That is, r_{0jk}, for example, defines the departures in each burst from the average RT. The residual terms are from a multivariate normal distribution with mean zero and two variance and one covariance term which represent the variance around the intercept ($\tau_{\pi 00}$) and around the slope ($\tau_{\pi 11}$) and the covariance between these two parameters ($\tau_{\pi 00}$).

At level 3, the grand mean of the intercept and slopes are defined by the γ parameters: γ_{000}, for example, defines the grand mean RT at the first session in the first burst. The individual departures from the grand mean are captured in the random effects *u*. Just as the random effects on level 2, the random effects at level 3 are from a multivariate normal distribution with mean zero and a variance-covariance matrix

$$\begin{bmatrix} \tau_{\beta 00} & \tau_{\beta 01} & \tau_{\beta 02} \\ \tau_{\beta 10} & \tau_{\beta 11} & \tau_{\beta 12} \\ \tau_{\beta 20} & \tau_{\beta 21} & \tau_{\beta 22} \end{bmatrix}. \quad (3)$$

Individual differences are captured in the variance estimates. The model defined in Equation (1) can be expanded by incorporating additional predictors, interaction terms, and, for example, quadratic slope terms.

As mentioned above, in the 3-level model defined in Equation (1) we separate WP and BP variance and locate its source on different levels with different timescales. In repeated measures designs the WP variability σ_ε^2, defined in Equation (2), is a source of variance typically treated as nuisance. However, one might be interested in changes of variability within persons. To illustrate the usefulness of such an approach, we will add predictors of WP variability (see Hedeker & Mermelstein, 2007) to examine changes in CRT variability from burst to burst and add a covariate to compare the variability in CRT of two subgroups based on a median split of the performance in the digit symbol test. That is, we introduce dummy-coded variables to identify participants who performed below the median ($DS_{group} = 0$) and participants who performed above the median ($DS_{group} = 1$) in the digit symbol test. Similarly, the four burst are coded from 0 to 3. By doing so, we specify that the error variance depends both on DS group membership and on the burst. Formally, the error variance is now weighted according to

$$\sigma_{\varepsilon jg}^2 = \sigma_\varepsilon^2 \exp(\xi_1 DS_{group} + \xi_2 Burst + \xi_3 BurstDS_{group}), \quad (4)$$

where ξ is the variance function coefficient and $\sigma_{\varepsilon jg}^2$ defines the error variance for group g = 0,1 and burst j = 0,1,2,3. In the first burst and for the DS group with the lower performance ($DS_{group} = 0$), the terms in the exponent are zero, which leaves exp(0) = 1 and $\sigma_{\varepsilon 00}^2 = \sigma_\varepsilon^2$. In the group that performed better in the DS task, the residual variance in the first burst is

$\sigma^2_{\varepsilon 01} = \sigma^2_{\varepsilon} \exp(\xi_1)$. The same logic applies to the burst and the interaction between the group membership and the burst. That is, depending on the value of ξ the residual variance σ^2_{ε} reduces, if $\xi < 0$ or increases if $\xi > 0$.

All models are estimated using R with the lme command from the nlme library (Pinheiro, Bates, DebRoy, Sarkar, & the R Development Core Team, 2011). The code for Model 3 is available in the Appendix. The significance of random effects are estimated using comparison of nested models with increasing numbers of random effects via the difference in −2*log-likelihoods associated to the restricted maximum likelihood (REML) estimate. The difference in a model where, for example, a random slope is estimated to a model without random slopes is in the variance of the slope and the covariance of the slope with the intercept. The latter model has two parameters less and, hence, both models can be compared via the likelihood ratio (LR) χ^2-test with 2 *df*.

Results from Project MIND

First we fit an unconditional model to obtain variance components at each of the three levels. These components can be related to the total variance using the intraclass correlation coefficient (ICC), which is a measure of variability between persons at Level 3 and between bursts and sessions at levels 2 and 1. The ICC for Level 3 was 0.45 (251484/(251484 + 27802 + 276457), at Level 2 it was ICC = 0.05, and at Level 1 it was ICC = 0.5. Apparently, most of the variance is at Level 1 and Level 3, indicating that the proportion of variance between persons is about 45% of the variance in the BRT. In turn, almost 50% of the variance is due to fluctuations within each person across sessions, and only 5% are attributed to variability within persons across bursts. This indicates that the performance in the CRT within the participants was relatively constant across bursts and the interest in explaining this little portion of the variance is usually not the primary concern. For purposes of this example, we still introduce variables at the burst level to explain that portion of the variance.

As can be seen in Table 1, the fixed effect γ_{000} indicated that, on average and across all participants, all bursts, and all sessions, the CRT was 1076 ms. The associated random effects indicated significant differences between persons, within bursts, and within sessions.

Next, in Model 2, we added slope terms defined by session and burst plus the quadratic terms of the slope variables in order to capture accelerations or decelerations in change over sessions and bursts. Given that these variables operate at different levels, they both have a different metric: Session is scaled in weeks while bursts are scaled in years. Further, we also included the crosslevel interaction between session and burst in order to account for changes in the session slopes over the years (i.e., bursts). As can be seen from Table 1, all fixed effects were statistically significant at the $p < .01$ level. The session slope indicated that every additional session decreased the CRT of the participants by, on average, 108 ms. At the same time, the quadratic slope term Session^2 indicated that across the sessions the gain in RT decelerated, and one can expect a benefit of about (3* − 108.3 + 9*17.6 = −166.5) 167 ms at the end of the four sessions in the first burst. Similarly, each burst lead to a reduction of approximately 147 ms – though again this reduction was nonlinear and was strongest from the first to the second burst and smallest for the burst in the fourth year. Further, the positive and significant ordinal interaction term indicated that the session slope becomes less steep as the years go by. This can be seen from Figure 1, which shows the average change (thick black line) across bursts surrounded by thin, gray lines, which represent individual predictions for the first 50 people of the sample.

The figure also underlines the large amount of individual differences in the CRT captured by the Level-3 random effects – denoted Person Level in Table 1. The random effects were

significant for both main effects but not for the crosslevel interaction term. That is, participants showed individual differences in the intercepts at the levels of the burst and the sessions. Further, the rate at which their reaction times decreased (for a few participants the RT increased) within and across bursts varied considerably from person to person. The correlations between the Level-3 random effects indicate that those who deviate positively from the average CRT intercept have steeper slopes both across sessions and across bursts. That is, larger initial RTs are associated with more change toward faster RTs. At the same time, the positive correlation between the session and burst slopes indicates that the rank order of change in the session slopes is largely maintained across bursts. In other words, participants with slow RTs benefited more from repeated practice both at the level of bursts and sessions. Moreover, both slopes were highly correlated, indicating that those who benefited most from repeated practice from session to session were almost the same as those who benefited from repeated practice from burst to burst.

At Level 2, across bursts, the random effects indicated intraindividual differences in the initial performance of bursts and change within burst across sessions. Both random effects indicated that participants deviated significantly from their average initial performance and their rate of change across sessions also varied considerably. Again, the negative correlation between the intercept and slope parameter indicated that slower RTs at the beginning of a burst are associated with a steeper slope, that is, more change across the sessions – and within the bursts.

In Model 3, we added four predictor variables which operate at different levels: Age, and $DS_{average}$ are both Level-3 between-person variables, whereas DS is a time-varying, within-person variable that operates at Level 2. DS scores may be different at each burst. Hence, these variables explain variance at different levels – at the between and within-person level. Given that a large portion of the total variance is located at the lowest level, Level-1, we also include variables that define the magnitude of the residual variance as described in Equation (4). That is, in order to capture changes in residual variability, we included burst and the dummy-coded DS_{group} as well as their interaction in the variance function. Note that these variables served as predictors for the variance function and captured changes in the residual variability across bursts and between two groups.

The parameter estimates from Model 3 are reported in Table 1. Note that the slope and squared slope coefficients are virtually unaltered compared to Model 2. In turn, the inclusion of the additional explanatory variables increased the intercept considerably to 2.3 s. The reason for this is $DS_{average,}$ which enters the model as an individually mean centered variable with an average of 42.6.

The effects on the RT of age at the first burst and $DS_{average}$ were both statistically significant. Note that these two variables are Level-3 variables, that is, they explain variability at the person level: Older adults were on average slower, with each year of additional age being associated with an increase of 17 ms in RT. In turn, the average performance in DS indicated that for each additional item that was correctly solved, the RT decreased by 18.9 ms. These two variables explained about 41% (ratio of Model 2 to Model 3 intercept variances: $(\gamma_{\beta 00M2} - \gamma_{\beta 00M3}/\gamma_{\beta 00M2})$ of the variance at the between-person level. At the burst level we introduced the grand mean centered variable DS which was assessed once per burst. The effect on the RT was statistically significant and indicated that every additionally remembered item decreases the RT by approximately 2 ms at each burst. Hence, the performance in DS accounted for differences in the RT intercept at each burst; in terms of explained variance, DS accounted for approximately 23% of the within-person variance at the burst level.

In order to model the residual variance, we introduced the burst and the DS_{group} variable. These parameters are defined by the ξ coefficients and are listed in the fixed effects part of Table 1. All three parameters, both main effects and the interaction term, were statistically significant. The residual variance changed according to the group and burst, the largest variance being at the first burst and for the dummy coded reference group $DS_{group} = 0$. That is, the estimate for the first burst and $DS_{group} = 0$ was $\sigma_{\varepsilon 00}=\sigma_{\varepsilon}^{2}\exp(0)=399318$, whereas the residual variance estimate for the group with better performance in DS was almost half as large, with $\sigma_{\varepsilon 01}=\sigma_{\varepsilon}^{2}\exp(-0.062*1)=206181$. Every additional burst decreased the variance for both groups by exp(−0.062*Burst). Further, the significant interaction term between group and burst indicated that for the group with the weaker performance in the DS task the decrease in the residual variability was somewhat larger. This relationship between group membership, burst, and residual variability is shown in Figure 2, where the upper line captures the predicted residual variability estimate for $DS_{group} = 0$ and the lower line shows the estimated variability for the participants who performed better in the DS task ($DS_{group} = 1$).

Discussion

The study of cognitive aging has long focused on several related questions that were addressed by a variety of longitudinal and cross-sectional designs. When does aging-related change in cognitive functioning begin? What is the impact of early life characteristics (e.g., childhood cognition) and contexts (e.g., parental SES, educational attainment) on later life cognition and change in cognition? What causes aging-related changes in cognition (i.e., due to underlying pathology /disease condition; inactivity) and what should be considered “normative”? Can these changes be prevented, delayed, or treated? It is increasingly being understood that the answers to these questions require a variety of longitudinal designs to capture the influences of early life on later life changes, key events, and lifestyle choices throughout the lifespan, and the identification of change-points that indicate an accelerated change relative to an individual’s characteristic level of performance. Attempts to answer these questions regarding the development and aging of cognitive functions across the lifespan has contributed to a rich array of substantive and methodological debates and advances.

These recent advances include application of innovative developmental designs, improvements in measurement for detecting within-person change, statistical advances in dynamic modeling and population inference conditional on mortality, and comparisons across birth cohorts and cultures. While we may not yet have definitive answers to many of these fundamental questions, the field of research on aging-related phenomena increasingly emphasizes the need for innovative longitudinal studies with recognition of the necessity of focusing on different temporal sampling of within-person change and variation. An increasing number of “measurement burst,” “daily diary,” and event-linked design studies (e.g., Moskowitz, Russell, Sadikaj, & Sutton, 2009) have recently been initiated. Such studies have the potential to improve the reliability of measurement and permit a better understanding of within-person dynamics.

Intensive Measurement Designs as Solutions for Questions of Gerontological Importance

Future research will be advanced by the use of intensive, but minimally obtrusive, measurement designs. Such designs permit the analysis of behavioral and biological processes as they unfold in both the short and long term and permit the opportunity to identify change from an individual’s normative performance (level and variation about characteristic level). One approach, described above, is the measurement burst design, which permits statistical decomposition of short-term variation and learning effects that overlay

normative aging and that provide a stronger basis for detecting accelerated change due to pathological processes. While the demands of such a design are generally much higher than typical longitudinal studies, such studies can be made more efficient by self-administered assessments performed in the participant's home, permitting intensive longitudinal data collection in large probability samples that are not constrained by geographical proximity to clinics and labs (e.g., Gallacher & Hofer, 2011). Internet-based surveys and testing may also serve to minimize problems associated with loss to follow-up and relocation and circumvent the problem of interval censoring that is an issue in widely spaced longitudinal designs. A central objective is to optimize the design and analysis features for particular research questions that require greater resolution of the processes of interest, and that permit reliable and sensitive assessment of functioning and change in functioning and of key determinants underlying short-term variation and long-term aging and health-related change.

Acknowledgments

Philippe Rast was supported by the Swiss National Science Foundation grant SNSF-131511. Stuart MacDonald was supported by a Career Investigator Scholar award from the Michael Smith Foundation for Health Research. Scott M. Hofer was supported by the NIH grant AG026453.

References

Bauer DJ. Evaluating individual differences in psychological processes. Current Directions in Psychological Science. 2011; 20:115–118.

Bryk, AS.; Raudenbush, SW. Hierarchical linear models. Newbury Park, CA: Sage; 1992.

Ferrer E, Salthouse TA, Stewart WF, Schwartz BS. Modeling age and retest processes in longitudinal studies of cognitive abilities. Psychology and Aging. 2004; 19:243–259. [PubMed: 15222818]

Fischer KW. Learning as the development of organized behavior. Journal of Structural Learning. 1980; 6:253–267.

Flynn JR. The mean IQ of Americans: Massive gains 1932 to 1978. Psychological Bulletin. 1984; 95:29–51.

Flynn JR. Massive IQ gains in 14 nations: What IQ tests really measure. Psychological Bulletin. 1987; 101:171–191.

Folstein MF, Folstein SE, McHugh PR. Mini-mental-state: A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research. 1975; 12:189–198. [PubMed: 1202204]

Gallacher J, Hofer SM. Generating large-scale longitudinal data resources for aging research. Journals of Gerontology: Psychological Sciences. 2011; 66B:i72–i79.

Hall CB, Derby C, LeValley A, Katz MJ, Verghese J, Lipton RB. Education delays accelerated decline on a memory test in persons who develop dementia. Neurology. 2007; 69:1657–1664. [PubMed: 17954781]

Hall CB, Lipton RB, Sliwinski MJ, Stewart WF. A change point model for estimating the onset of cognitive decline in preclinical alzheimer's disease. Statistics in Medicine. 2000; 19:1555–1566. [PubMed: 10844718]

Hedeker, D.; Mermelstein, RJ. Mixed-effects regression models with heterogeneous variance: Analyzing Ecological Momentary Assessment (EMA) data of smoking. In: Little, TD.; Bovaird, JA.; Card, NA., editors. Modeling contextual effects in longitudinal studies. Mahwah, NJ: Erlbaum; 2007. p. 183-206.

Hofer SM, Flaherty BP, Hoffman L. Cross-sectional analysis of time-dependent data: Problems of mean-induced association in age-heterogeneous samples and an alternative method based on sequential narrow age-cohorts. Multivariate Behavioral Research. 2006; 41:165–187.

Hofer, SM.; Rast, P.; Piccinin, AM. Methodological issues in research on adult development and aging. In: Whitbourne, SK.; Sliwinski, MJ., editors. Handbook of developmental psychology: Adult development and aging. New York: Wiley-Blackwell; 2012. p. 72-93.

Hofer SM, Sliwinski MJ. Understanding aging: An evaluation of research designs for assessing the interdependence of aging-related changes. Gerontology. 2001; 47:341–352. [PubMed: 11721149]

Hoffman L, Hofer SM, Sliwinski MJ. On the confounds among retest gains and age-cohort differences in the estimation of within-person change in longitudinal studies: A simulation study. Psychology and Aging. 2011; 26:778–791. [PubMed: 21639642]

Hultsch DF. Learning to learn in adulthood. Journal of Gerontology. 1974; 29:302–308. [PubMed: 4821845]

Hultsch DF, MacDonald SWS, Dixon RA. Variability in reaction time performance of younger and older adults. Journal of Gerontology: Psychological Sciences. 2002; 57B:101–115.

Hultsch, DF.; Strauss, E.; Hunter, MA.; MacDonald, SWS. Intraindividual variability, cognition, and aging. In: Craik, FIM.; Salthouse, TA., editors. The handbook of aging and cognition. 3. New York: Psychology Press; 2008. p. 491-556.

Jacqmin-Gadda H, Commenges D, Dartigues JF. Random changepoint model for joint modeling of cognitive decline and dementia. Biometrics. 2006; 62:254–260. [PubMed: 16542253]

Kraemer HC, Yesavage JA, Taylor JL, Kupfer D. How can we learn about developmental processes from cross-sectional studies, or can we? American Journal of Psychiatry. 2000; 157:163–171. [PubMed: 10671382]

Kurland BF, Johnson LL, Egleston BL, Diehr PH. Longitudinal data with follow-up truncated by death: Match the analysis method to research aims. Statistical Science. 2009; 24:211–222. [PubMed: 20119502]

Lindenberger U, Baltes PB. Testing-the-limits and experimental simulation: Two methods to explicate the role of learning in development. Human Development. 1995; 38:349–360.

Lövdén M, Li SC, Shing YL, Lindenberger U. Within-person trial-to-trial variability precedes and predicts cognitive decline in old and very old age: Longitudinal data from the Berlin Aging Study. Neuropsychologia. 2007; 45:2827–2838. [PubMed: 17575988]

MacDonald SWS, Hultsch DF, Dixon RA. Performance variability is related to change in cognition: Evidence from the Victoria Longitudinal Study. Psychology and Aging. 2003; 18:510–523. [PubMed: 14518812]

Martin M, Hofer SM. Intraindividual variability, change, and aging: Conceptual and analytical issues. Gerontology. 2004; 50:7–11. [PubMed: 14654720]

Molenaar, PCM.; Huizenga, HM.; Nesselroade, JR. The relationship between the structure of interindividual and intraindividual variability: A theoretical and empirical vindication of developmental systems theory. In: Staudinger, UM.; Lindenberger, U., editors. Understanding human development: Dialogs with lifespan psychology. Netherlands: Kluwer; 2003. p. 339-360.

Moskowitz DS, Russell JJ, Sadikaj G, Sutton R. Measuring people intensively. Canadian Psychology/ Psychologie Canadienne. 2009; 50:131–140.

Nesselroade, JR. Interindividual differences in intraindividual change. In: Collins, LM.; Horn, JL., editors. Best methods for the analysis of change. Washington, DC: American Psychological Association; 1991a. p. 92-105.

Nesselroade, JR. The warp and the woof of the developmental fabric. In: Downs, R.; Liben, L.; Palermo, DS., editors. Visions of esthetics, the environment, and development: The legacy of Joachim F Wohwill. Hillsdale, NJ: Erlbaum; 1991b. p. 213-240.

Nesselroade JR, McCollam KMS. Putting the process in developmental processes. International Journal of Behavioral Development. 2000; 24:295–300.

Newell KM, Liu YT, Mayer-Kress G. Time scales in motor learning and development. Psychological Review. 2001; 108:57–82. [PubMed: 11212633]

Pinheiro, J.; Bates, D.; DebRoy, S.; Sarkar, D. the R Development Core Team. Computer software manual. Vienna, Austria: R Foundation for Statistical Computing; 2011. nlme: Linear and nonlinear mixed effects models. R package version 3.1–102.

Rabbitt P. Does it all go together when it goes? The nineteenth Bartlett memorial lecture. Quarterly Journal of Experimental Psychology. 1993; 46A:385–434. [PubMed: 8378549]

Ram N, Gerstorf D. Time-structured and net intraindividual variability: Tools for examining the development of dynamic characteristics and processes. Psychology and Aging. 2009; 24:778–791. [PubMed: 20025395]

Rast P, Hofer S, Sparks C. Modeling individual differences in within-person variation of negative and positive affect in a mixed effects location scale model using BUGS/JAGS. Multivariate Behavioral Research. 2012; 47:177–200.

Raudenbush, SW.; Bryk, AS. Hierarchical linear models: Applications and data analysis methods. Vol. 1. Thousand Oaks, CA: Sage, Inc; 2002.

Rönnlund M, Nilsson LG. The magnitude, generality, and determinants of flynn effects on forms of declarative memory and visuospatial ability: Time-sequential analyses of data from a Swedish cohort study. Intelligence. 2008; 36:192–209.

Salthouse TA. When does age-related cognitive decline begin? Neurobiology of Aging. 2009; 30:507–514. [PubMed: 19231028]

Salthouse TA. Effects of age on time-dependent cognitive change. Psychological Science. 2011; 22:682–688. [PubMed: 21467547]

Schaie, KW. Intellectual development in adulthood: The Seattle Longitudinal Study. New York: Cambridge University Press; 1996.

Schaie, KW. Historical processes and patterns of cognitive aging. In: Hofer, SM.; Alwin, DF., editors. Handbook of cognitive aging: Interdisciplinary perspectives. Thousand Oaks, CA: Sage; 2008. p. 368-383.

Sliwinski MJ. Measurement-burst designs for social health research. Social and Personality Psychology Compass. 2008; 2:245–261.

Sliwinski, MJ.; Hoffman, L.; Hofer, SM. Modeling retest and aging effects in a measurement burst design. In: Molenaar, PCM.; Newell, KM., editors. Individual pathways of change: Statistical models for analyzing learning and development. Washington, DC: American Psychological Association; 2010. p. 37-50.

Sliwinski MJ, Smyth JM, Hofer SM, Stawski RS. Intraindividual coupling of daily stress and cognition. Psychology and Aging. 2006; 21:545–557. [PubMed: 16953716]

Thorvaldsson V, Hofer SM, Berg S, Johansson B. Effects of repeated testing in a longitudinal age-homogeneous study of cognitive aging. Journal of Gerontology: Psychological Sciences. 2006; 61B:P348–P354.

Walls, TA.; Barta, WD.; Stawski, RS.; Collyer, C.; Hofer, SM. Timescale-dependent longitudinal designs. In: Laursen, B.; Little, TD.; Card, N., editors. Handbook of developmental research methods. New York: Guilford; 2011. p. 46-64.

Wechsler, D. WAIS-R: Wechsler Adult Intelligence Scale Revised. New York: Harcourt, Brace, & Jovanovich; 1981.

Wohlwill, JF. The study of behavioral development. New York: Academic Press; 1973.

Yan Z, Fischer K. Always under construction – dynamic variations in adult cognitive microdevelopment. Human Development. 2002; 45:141–160.

Zimprich D, Rast P. Verbal learning changes in older adults across 18 months. Aging, Neuropsychology, and Cognition. 2009; 16:461–484.

Appendix

R Code for Model 3

```
library(nlme)

# load nlme library

## Session2 = Session*Session

## Burst2 = Burst*Burst

## DS.c = Digit Symbol grand-mean centered

## Age.c = Age of respondents grand-mean centered
```

```
## DS.pc = Digit Symbol person-mean centered

## DS.grp = Dummy coded DS variable via median split

## Subject = Subject identifier

model3 ##← lme(CRT ~ 1 + Session*Burst + Session2 + Burst2 + DS.c + Age.c + DS.pc,

    # fixed effects are defined

    random = list(Subject = ~ 1 + Burst + Session, Burst = ~1 + Session),

    # random effects are defined. Note the hierarchy

    # where Burst is nested in Subject using list()

    weights = varComb(varExp(form = ~DS.grp), varExp(form = ~Burst), varExp(form =
    ~I(DS.grp*Burst))),

    # Variance function is given here (see Equation 4):

    # each effect plus the interaction term in I() need

    # to be defined using varComb()

    control = lmeControl(msVerbose = T, maxIter = 150, msMaxIter = 80),

    # some control arguments are passed along – they are not

    # necessary and may be omitted.

    data = crt.MIND)

    # data statement: Data needs to be in long format.

    # Consider eg. make.univ() from library(Multilevel)

summary(model3)
```

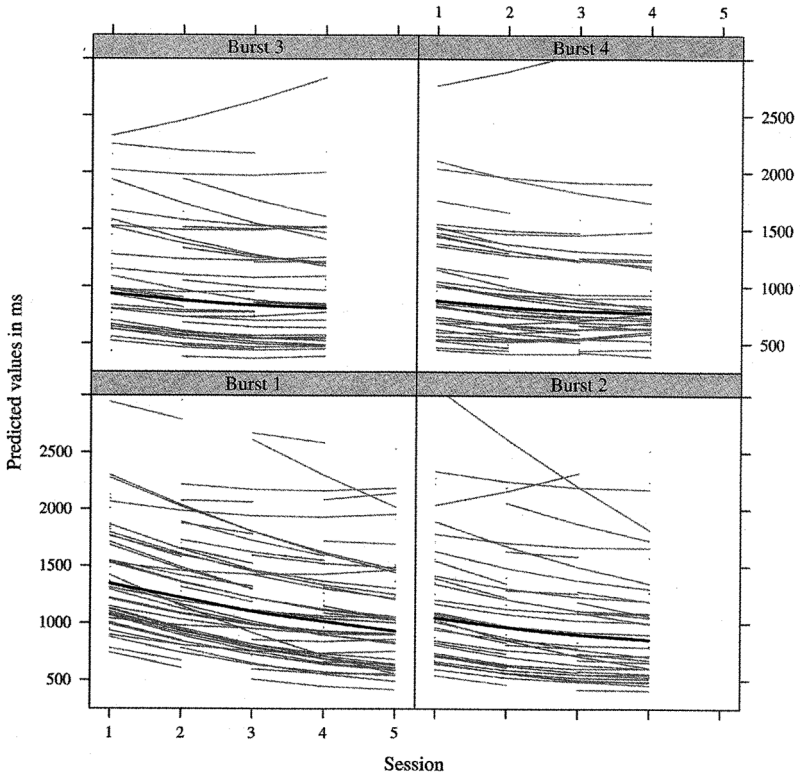

Figure 1.

Fitted trajectories from Model 2. The thick black lines are average effects, and the thin gray lines represent individual trajectories. Depicted are the first 50 participants. Note that the first burst consisted of five sessions and the consecutive bursts consisted of four sessions.

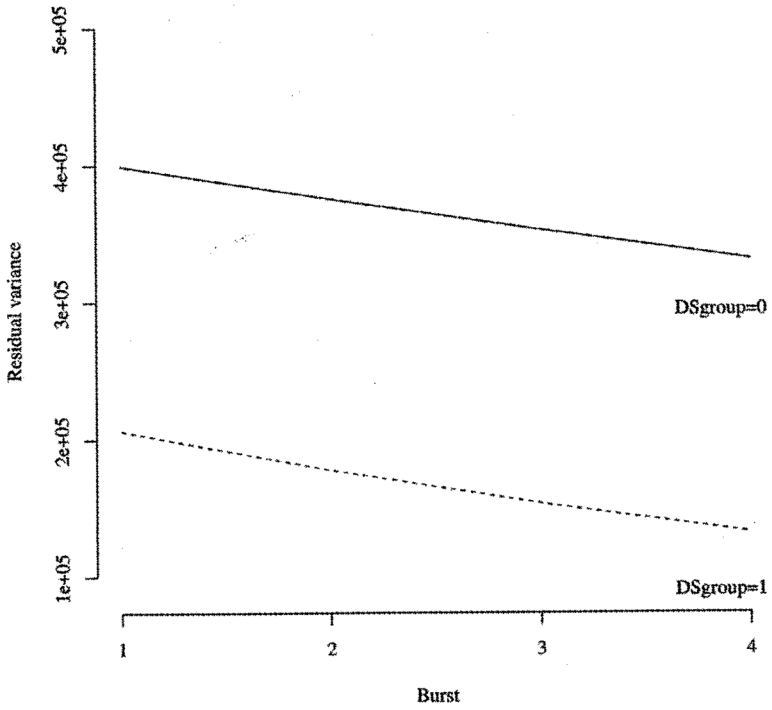

Figure 2.
Predicted residual variability for two groups defined by a median split of DS. DSgroup = 0 are participants who performed lower than the median of 43 and DSgroup = 1 are participants who scored 43 or more points in the DS task.

Table 1

Three-level analysis of CRT data

	Model 1			**Model 2**			**Model 3**		
Fixed effect	**Estimate**	***SE***	***p***	**Estimate**	***SE***	***p***	**Estimate**	***SE***	***p***
Intercept	1075.9	29.4	<01	1305.3	34.6	<.01	2112.4	107.4	<.01
Session				−108.3	3.8	<.01	−107.3	3.7	<.01
Burst				−147.2	6.4	<.01	−163.	7.4	<.01
$Session^2$				17.6	0.8	<.01	16.4	0.8	<.01
$Burst^2$				38.8	3.3	<.01	38.4	3.2	<.01
Session*Burst				31.8	1.7	<01	31.1	1.7	<.01
DS							−2.3	0.8	<.01
Age							16.9	3.9	<.01
$DS_{average}$							−18.9	2.4	<.01
Variance function parameters							Coefficient	$LR\chi^2$	*p*
$\xi_{DSgroup}$							−0.661	5501	<01
ξ_{Burst}							−0.062	265	<01
$\xi_{DSgroup*Burst}$							−0.086	288	<.01
Random effect	Estimate	$LR\chi^2$	*p*	Estimate	$LR\chi^2$	*p*	Estimate	$LR\chi^2$	*p*
Person Level									
Intercept	251484	128066	<.01	335596	1472	<.01	197952		
Burst				4640	191	<.01	5178	185	<.01
Session				1597	206	<.01	1499	171	<.01
$r_{I,B}$				−.51			−.56		
$r_{I,S}$				−.79			−.70		
$r_{B,S}$				.85			.82		
Burst Level									
Intercept	27802	13838	<.01	28658	2145	<.01	21886	2054	<.01
Session				2958	1625	<.01	2959	1692	<01
$r_{I,S}$				−.75			−.77		
Session Level									

Fixed effect	Model 1			Model 2			Model 3		
	Estimate	*SE*	*p*	Estimate	*SE*	*p*	Estimate	*SE*	*p*
Residual	276457			258632			399318		

Note. The LRχ^2 is the difference in the −2*log-likelihood of a model where the respective random effect is omitted.