How Stable are Self-Reported Measures of Media Use?

Jacob A. Long

Contact: jacob.long@sc.edu

Abstract

Measuring media exposure is challenging due to apparently unreliable self-reports. As a result, media exposure appears highly stable after correcting for measurement error. The model commonly used to quantify reliability and stability makes untenable assumptions about the data used in communication. Commonly violated assumptions can bias reliability estimates downward and stability estimates upward. Despite high reported stability, media use changes are associated with theoretically expected outcomes, suggesting more meaningful variation than captured by current methods.

How Stable are Self-Reported Measures of Media Use?

Valid and reliable measurement of media exposure is a major challenge for mass communication. Although the particulars are debatable, it is clear that the most common approach to measurement — self reports on questionnaires — have significant problems with accuracy (Guess, 2015; LaCour & Vavreck, 2014; Prior, 2013). Technological advances have made passive measurement sufficiently feasible to occasionally compare questionnaire measures of media consumption to observed consumption and such comparisons invariably show self-reports to be quite inaccurate. The question that is debated is whether those self-reported measures are too wrong to be useful for some of the research questions they are used to pursue. Consistent over-reporting, for instance, does not render between- or within-subjects comparisons invalid even if the mean estimates of media use are wrong. More problematic would be if the accuracy or direction of inaccuracy is related to other key variables, but research has not yet demonstrated that this could be problematic for inference; for instance, responses to news use measurements do not appear to be related to the propensity to give socially desirable responses (Eveland et al., 2009). Jürgens, Stark, and Magin (2019) point out that methods for passive tracking are hardly fool-proof as well, inducing bias in the samples available for this kind of tracking as well as documenting what appeared to be a technical error in a commercially available tool for media tracking that would only have been apparent on close analysis. There are simply no perfect options for media measurement. Although my focus will be specifically on political media use, especially partisan media, I will discuss what is known about media use more broadly in part because the issues are fairly general.

This unreliability (to say the least) in measurement was captured well in a recent meta-analysis by Scharkow (2019). The estimates of reliability and stability from that study led its author to question the use of repeated self-reported measures of media exposure at all in longitudinal research, given the apparent lack of change in the underlying variable after correction for measurement error, which appears very high. As will be further explained, with the available data, measurement error and stability are in tension. The greater the measurement error, the more we typically assume the underlying construct is stable. Likewise, as Scharkow notes, if media use is indeed extremely stable, it calls into question both our collective, practical ability to study its effects (which require variation to observe) and therefore the validity of the field’s many theories that assume media use is variable over time. Besides effects of media, any theoretical approach that treats media as an outcome — like selective exposure theories — face the same challenge if most people do not experience much change over time in their media use. In this paper, I will review some of the evidence on the measurement of media use, the stability of media use, and then discuss in detail the statistical issues that make it difficult to know whether its stability is really so high as to almost preclude using self-reported measures for mass communication research.

Media Measurement Challenges and Observed Stability of Media Use

One complication is sorting out what exactly is being measured. Recent developments in measurement have addressed overly broad measures used in the past that made it difficult to determine what kinds of content people were exposed to. The “program list technique” (Dilliplane et al., 2013) — which was designed for television but is just as easily ported to print, digital, and radio sources — presents respondents with a list of media sources. They are then asked if, over the course of a specified timeframe (usually a week or month), they have used the source regularly. The measure follows best practices identified by research on questionnaire design: reports of behavior are more accurate when broken into their constituent parts (Menon, 1997), granular response choices can influence responses in difficult-to-predict ways (Schwarz et al., 1985), and regular behaviors are reported most accurately (Menon, 1993). Despite concerns that the number of programs does not correspond with amount of exposure (Prior, 2013), validation research based on a sophisticated passive tracking system found a close relationship between the number of programs and total amount of media consumption (LaCour & Vavreck, 2014). The decomposition into separate programs also suits a wider variety of research problems, allowing for distinctions to be made about the kind, not just degree, of media consumption. A common use of this measure has been to differentiate consumption of partisan media sources from each and from non-partisan sources (Dilliplane, 2014; Long et al., 2019).

Given the existence of some measures suited to assess the stability of media use, how stable is it? Evidence suggests it is very stable, at least over the course of time that a typical panel study covers. In roughly a 6-month timeframe, Dilliplane et al. (2013) found a very high stability coefficient of over 0.9 for the program list technique using television programs. Nevertheless, in a subsequent study Dilliplane (2014) found that changes, not absolute levels, in media use were most influential in changing vote choices and affect towards candidates in the 2008 election. This suggests the coarseness of the program list measure may not capture some change in media consumption until it becomes fairly significant or, perhaps, the Heise (1969) stability measure overstates the proportion of variation that should be attributed to measurement error — a possibility I will soon explore. Scharkow (2019) argues Heise's stability measure is inappropriately applied to the list technique as Heise assumes an interval measure whereas the list technique produces a sum of binary indicators. Nonetheless, a meta-analysis of other measures of television, internet, and print media (primarily news) used in 33 3-wave panel studies — mostly using hours/day or days/week frequency measurements — found an estimated stability coefficient of 0.9, quite consistent with Dilliplane et al.'s (2013) findings (Scharkow, 2019).

Although none of these studies cover the timespan that studies of partisan identification do (e.g., Alwin & Krosnick, 1991), the available evidence is at least suggestive that media use may be even more stable than party affiliation. This comes as a surprising finding if we assume it is true. There are some reasons, beyond prior expectations, to be skeptical of this interpretation. First, change in media use and change in social identity are being operationalized quite differently in the studies I have reviewed. Variation in party affiliation, as it has been studied by most political scientists, necessarily involves categorical change. It would be unusual to treat media use this way. One can conceptualize multiple categories of media use, but they are rarely mutually exclusive and usually vary in both degree and kind. Further, the kind of stability captured by the Heise (1969) measure concerns rank-order. In other words, the high stability of media usage measured by Dilliplane et al. (2013) and meta-analyzed by Scharkow (2019) is best interpreted as showing that it is unusual for someone among the lightest media users to become among the heaviest media consumers and vice versa. Intra-individual change is basically ignored by this measure except when it changes the rank ordering of respondents. These empirical results are also consistent with a model of media use varying in relatively short timeframes after which it reverts back to its mean.

This is important because intra-individual change in media use — a person differing from their norms — is quite often associated with theoretically expected changes in other variables. Fixed effects panel models, which completely disaggregate effects of changes in variables in a way that is robust to confounding (Allison, 2009), have shown changes in political media use can increase polarization (Dilliplane, 2014; Garrett et al., 2014; Wojcieszak et al., 2017), political participation (Kruikemeier & Shehata, 2017; Shehata et al., 2016), political interest (Kruikemeier & Shehata, 2017; Moeller et al., 2018), and political expression (Gervais, 2014). When a large portion of the change in a longitudinal variable is due to measurement or some other kind of random error, then most of the variation available to model with fixed effects regression would be measurement error.

The impressive pattern of results in these models that rely on intra-individual variability for statistical power in communication research suggest there may be more than meets the eye when it comes to reports of high stability in media use. As I have suggested and will explore in more detail later, the quantification strategies may overlook the more meaningful sort of variation. But if we accept the findings of high stability at face value, it suggests something that is perhaps not obvious. Fairly substantial effects as a result of changes in media use are observed despite its high stability, at least as stability is usually quantified. This is consistent with media being highly influential; that it is not more obvious owes only to the relative infrequency with which media habits change. Of course, this is just one interpretation. It is arguably tautological to claim the variation is important (or greater than claimed) based solely on positive statistical results. More granularity is also possible by analyzing the content of media (as suggested by Scharkow, 2019), which is likely to change more than media consumers change their selections (de Vreese et al., 2017), or by zooming in on day-to-day variations.

The stability of a construct refers to how little it changes. Something that is completely stable does not change at all. Besides constructs that are immutable, we can treat stability as a property that varies continuously; some constructs are more (un)stable than others. It is not possible to engage in a nuanced discussion of stability without dealing with both operationalization and conceptualization. In the coming pages, my discussion of stability will be rather technical, but primarily in the service of having a better conceptual foundation. I begin with a simple description and then move on to a detailed explanation of the prevailing methods for measuring stability, which originate in research on measurement reliability. I show how these approaches have rarely-discussed assumptions that may lead to incorrect conclusions about the stability of constructs. Part of the reason for this detailed discussion is because a core threat to the validity of many popular approaches to mass communication research is the possibility that the constructs under study are simply too stable to be studied. Grasping these issues will be needed to undertake the necessary research to sort out whether the stability is indeed just too high.

Quantifying Stability of Constructs

Typically, to say a construct is stable is to put it in contrast to some reference standard. Such a standard is not often explicitly invoked, but the description “stable” implies the suitability of some types of analytic and theoretical approaches and the exclusion of others. For instance, in the analysis of repeated measures data, stable variables are considered a challenge because they do not exhibit enough within-unit variance to have statistical power (e.g., Clark & Linzer, 2015). Such claims are sensitive to the timeframe under consideration. As will be discussed in more detail later, media use is considered a highly stable behavior; yet in the course of a single day, the typical person will stop and start media use many times. It is only when aggregating at daily, weekly, or other levels that stability appears. Which treatment of time is appropriate depends on the research setting.

Assessing the stability of a construct requires measuring it repeatedly on the same subjects[1] over time. Armed with multiple measurements, the extent to which a subject's repeated measurements differ from one another is an indicator of stability. In applied social scientific research, however, one typically must distinguish measurement error from actual changes in the underlying construct. This assumes that the measurement $\mathrm{X_t}$ at time t is the sum of the true value T_t and random measurement error e_t:

$$X_t = T_t + e_t$$

This, as with most technical research on stability, comes from psychometric research on reliability (Lord & Novick, 1968). To assess reliability, the observed variance of X_t must be partitioned into random measurement error variance and true score variance:

$$Var(X_t) = Var(T_t) + Var(e_t)$$

The same information is needed to assess stability. It follows that

$$Var(T_t) \leq Var(X_t)$$

Although measurement and study design may aggregate out variance that would be observed at different time intervals (or timespans) as discussed earlier, a study design cannot support the

[1] I will use the language of human subjects research, but this largely applies to other units of analysis.

estimation of more variance[2] in T_t than there is observed variance in X_t. As a logical extension of these relationships, any overestimate of error variance e_t will also overestimate stability. The reliability, denoted α^2, is defined in this framework, as

$$\alpha^2 = \rho_{XT}^2 = \frac{Var(T)}{Var(X)} = \frac{Var(T)}{Var(T) + Var(e)}$$

In plain terms, reliability is the proportion of observed variance that is due to variance in the underlying construct. The alternate notation of ρ_{XT}^2 signifies that reliability can be conceptualized as the (squared) correlation between the observed and true values. The more reliable the measurement X_t is, the more similar $Var(X_t)$ and $Var(T_t)$ will be. If reliability is underestimated, stability is overestimated. I point this out because much more research focuses on the proper estimation of reliability than there is research explicitly dealing with stability. One exception in which the reliability approach was used in a debate about construct stability was in political science, in which it has been argued that what was once described as incoherence (or non-existence) of political attitudes in the general public (Converse, 1962, 1964) was actually an artifact of highly unreliable measurements that are more consistent with quite stable attitudes once measurement error is corrected (Achen, 1975; Green & Palmquist, 1990).

---

[2] This is not strictly true. In a "Heywood case" (Heywood, 1931), the reliability may be estimated to be greater than 1 (e.g., the measure explains more than 100% of variance of the true scores) and measurement error variance is therefore negative. Because these estimates cannot correspond with reality, I do not include them as estimates that could be treated as credible.

To have enough information to statistically estimate $Var(T_t)$ and $Var(e_t)$, at least three t are required per subject (Heise, 1969). From here forward I will discuss these quantities as if they were obtained in a three-wave panel, but the approach generalizes to any greater number of measurement periods. Figure 2 shows a theoretical and empirical model connecting these concepts, which Alwin (2007, p. 105) calls the “quasi-simplex model” but appears without a name in other parts of the literature (Heise, 1969; Wheaton et al., 1977; D. E. Wiley & Wiley, 1970).

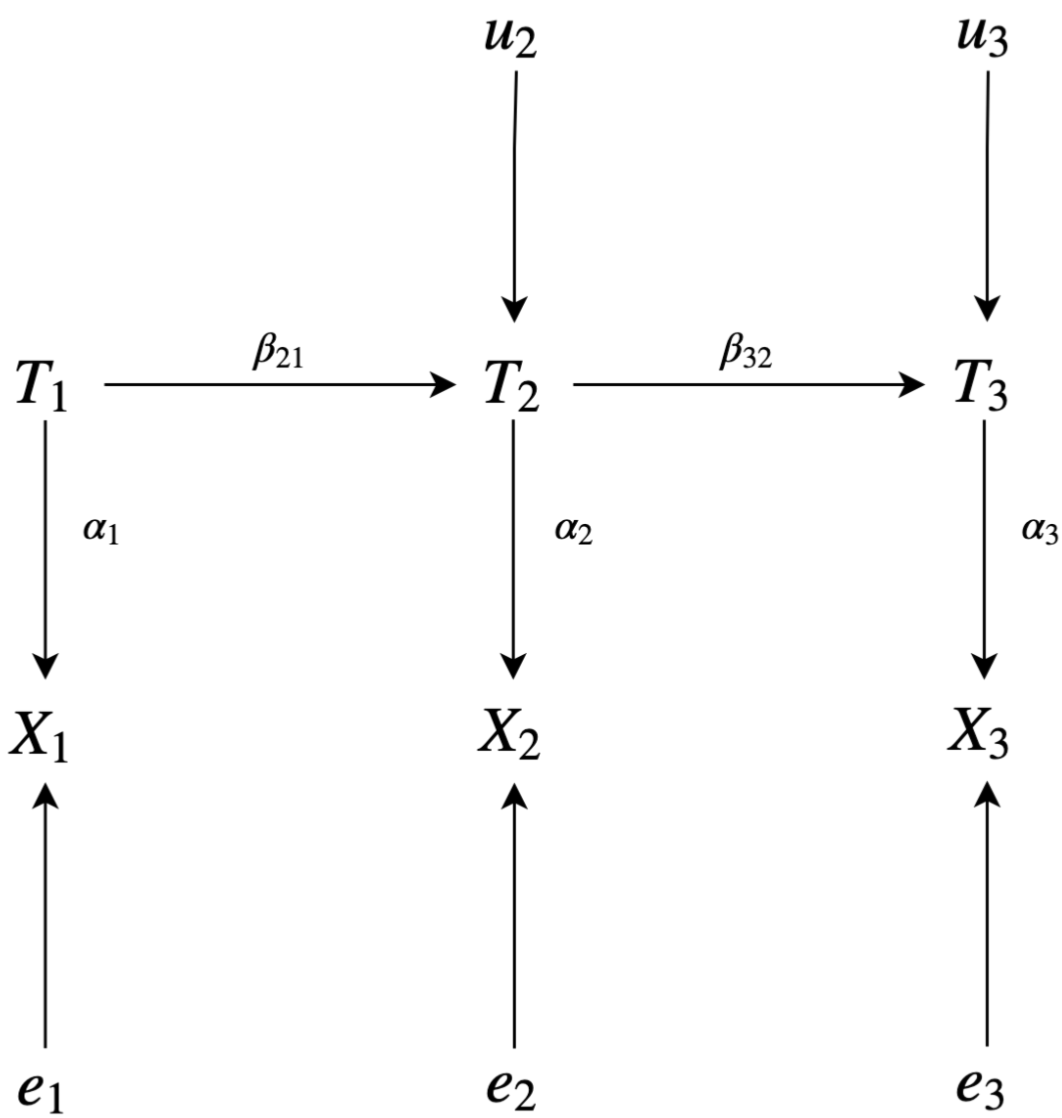

Figure 1: Quasi-simplex Model

In the quasi-simplex model, the true score at time t is understood to be determined by the true score at time $t-1$ as well as true score variance, $Var(T_t) = u_t$:

$$T_t = \beta_{t,t-1} T_{t-1} + u_t$$

$\beta_{t,t-1}$ represents one type of stability measurement. In the three-wave quasi-simplex model, β_{21} and β_{32} (see Figure 1) are two separate estimates of wave-to-wave stability. When standardized, $\beta_{t,t-1}$ is interpreted as the correlation between true scores from one wave to the next. The quasi-simplex model explains the observed correlation $r_{t,t-1}$ as:

$$r_{t,t-1} = \rho_{XT}^2 \beta_{t,t-1}$$

In other words, the observed correlation is the correlation between true scores as attenuated by reliability. This means $r_{t,t-1} \leq \beta_{t,t-1}$ except in Heywood cases. As my discussion proceeds, it may be useful to keep in mind that the practical boundaries for the true value of $\beta_{t,t-1}$ are $r_{t,t-1}$ at the lower bound and 1 at the upper bound. When I raise issues that suggest $\beta_{t,t-1}$ is often estimated in a way that is biased upward, the reader should bear in mind the true value cannot be lower than $r_{t,t-1}$.

The quasi-simplex model embeds several assumptions (see Alwin, 2007; Wheaton et al., 1977):

1. T_t is only influenced by its value at $t-1$ (*lag-1 assumption*)

2. There is no serial correlation in the measurement errors (i.e., $Cov(e_t, e_{t-p}) = 0$ for all p in 1 ... P where P is the number of panel waves prior to t)

3. True scores and measurement error are uncorrelated (i.e., $Cov(T_t, e_t) = 0$).

4. True score variance is not correlated with prior true scores (i.e., $Cov(u_t, T_{t-p}) = 0$ for all p in 1 ... P where P is the number of panel waves prior to t)

In the Heise (1969) quasi-simplex model, which is most often used by applied researchers to quantify stability, the reliability coefficients α_t are also constrained to be equal over time in order to identify the model[3]. A stability coefficient is estimated using equations in Heise (1969) that standardize the $\beta_{t,t-1}$ parameters to be interpretable as correlations and in the three-wave case, they are often multiplied together to yield a coefficient equivalent conceptually to β_{31}. This quantity, which can be computed using observed correlations, is fundamentally a measure of the rank order stability of subjects over time. That means if all subjects shift in the same direction and at the same magnitude over time, there is still absolute stability. An example of this is age — although age goes up as measurements are repeated, nobody becomes older or younger than anyone else relative to the first measurement. This alone would not affect stability estimates using the quasi-simplex model. Whether this captures the kind of stability one is interested in depends on the research question, but it is the primary if not sole measure of stability used in

[3] D.E. Wiley and J.A. Wiley (1970) instead constrain the error variances e_t to be equal, meaning the reliability can vary across time. These are two different assumptions that are not sufficiently important to explore in detail here.

communication research (see Lee et al., 2008; Scharkow, 2019). For the time being, I take this approach on its own terms as a valid way to treat stability. There is also a case in which the interpretation of rank order-stability more closely matches the intuition of stability as lack of change. If the aggregate time series is stationary — that is, the aggregated mean and variance are constant over time, which is not extraordinary — then one can relatively safely interpret rank-order stability to mean lack of change.

A useful way to interpret the stability estimates $\beta_{t,t-1}$ (when standardized) is to consider their implications for analysis of the construct under study. Take for example a dataset in which the observed correlation of X from time 1 to 2 (r_{21}) and 2 to 3 (r_{32}) is .65 and the correlation from 1 to 3 (r_{31}) is .60. Using the Heise (1969) formulas, the reliability would be:

$$\rho_{XT}^2 = \frac{r_{21}r_{32}}{r_{31}} = \frac{.65 \times .65}{.6} \approx .70$$

Stability parameters would be:

$$\beta_{21} = \frac{r_{31}}{r_{32}} \approx 0.92, \qquad \beta_{32} = \frac{r_{31}}{r_{21}} \approx 0.92, \qquad \beta_{31} = \frac{r_{31}^2}{r_{21}r_{32}} \approx 0.85$$

Panel data regression models that focus on intraindividual change, like fixed effects regression and many offshoots (Allison, 2009), discard all between-subjects variance by design. If there is a high $\beta_{t,t-1}$ like 0.92, then only about 15% (i.e., $1 - 0.92^2$) of $Var(T_t)$ can be modeled. Whether this is a theoretically interesting amount of variance that is "up for grabs," so to speak, depends on the magnitude of $Var(T_t)$. Since T_t is not observed, we ultimately are modeling X_t. Given $r_{t,t-1}$ is 0.65, around 58% of the variation of X_t is subject to analysis but if the estimated reliability is correct, much of that variance is measurement error.

To show the consequences of underestimated reliability, let us suppose reliability is actually 0.9. We can then use the equality described earlier, $r_{t,t-1} = \rho^2_{XT}\beta_{t,t-1}$, and solve for $\beta_{t,t-1}$:

$$\beta_{21} = \beta_{32} = \frac{r_{32}}{\rho^2_{32}} = \frac{.65}{.9} \approx .72$$

Although both 0.72 and 0.92 are reflective of fairly stable variables, the former figure is far more manageable from the perspective of statistical power (see Clark & Linzer, 2015).

The quasi-simplex model raises several theoretical questions. One is an assumption besides those I have already enumerated about this model: There are no other causes of T_t besides T_{t-1}, and if we relax the lag-1 assumption, still no other causes other than $T_{p<t}$. An alternate model, visualized in Figure 3, allows for another cause of T_t. In this model, adapted from Wiley and Wiley (1970), Y is a variable that also affects the value of T_t. For simplicity, Y is a time-invariant construct but it may theoretically also be something that changes across periods. Few are so naïve as to think that interesting social scientific constructs are actually caused only by their values in the past — some claim they rarely are caused by past values whatsoever (Achen, 2000) — so the question about how to specify the model again comes down to a mixture of empirical and theoretical considerations.

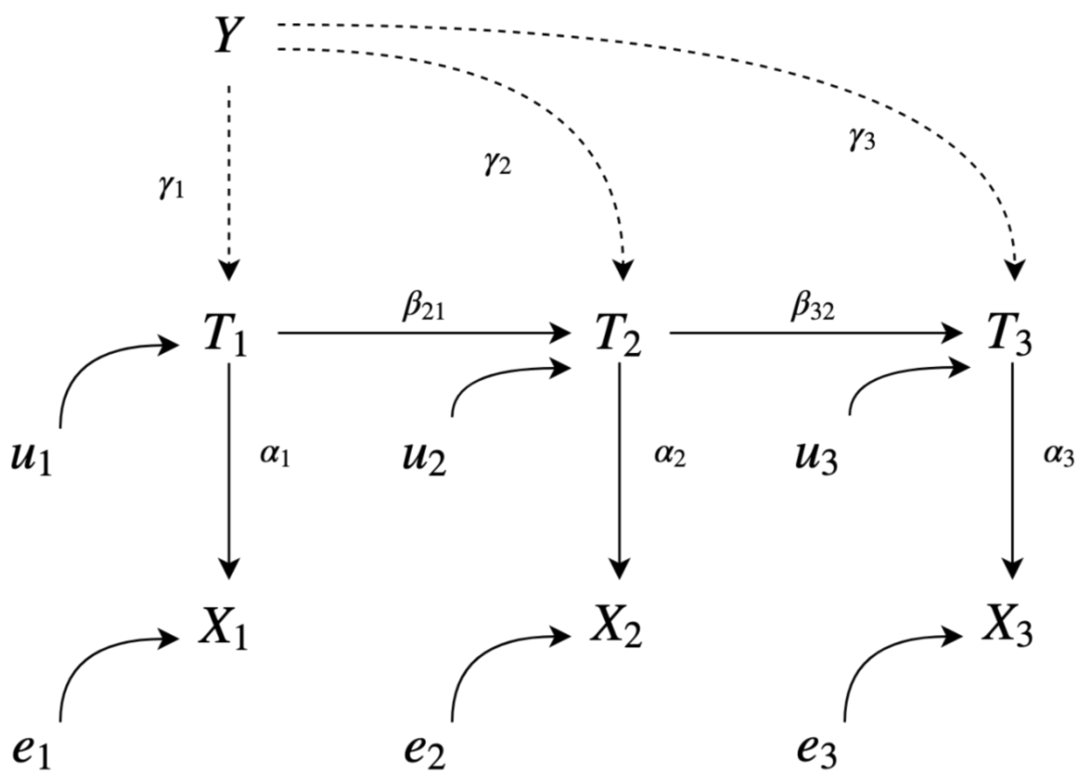

Figure 2: Stability Model with an Unmeasured Second Variable

In this case, the question is: what is stability? The quasi-simplex model and others like it are most useful when the interest is in identifying $Cov(T_t, T_{t-1})$ but there is no interest in u_t, the variance in T_t that is not caused by past values. This can be a valid simplifying assumption to make when the goal is simply characterizing stability while remaining agnostic about its causes. It is not an objective measure of $Cov(T_t, T_{t-1})$ because T_t is not directly observed — but it is a relatively simple approximation. Stability in this empirical-theoretical framework is what the over-time correlation of successive measures of T_t would be *if they were measured without error*. What is useful about a more fully-specified model, which will generally shrink the $\beta_{t,t-1}$ terms, is engaging in more complex counterfactual thinking. Using the multi-variable model in Figure 2, the interpretation of the $B_{t,t-1}$ is what the over-time covariance of successive measures of T_t would be if they were measured without error *and the other causal variable Y was removed* *from the system*. This allows for T_t to appear stable but not be inherently stable. In other words, fully-specified models may lead to the conclusion that T_t is only stable due to the causal influence of other factors and would presumably not be stable if those causes were somehow removed or changed. This is an argument that can be made about the relationships between communication and psychological variables (Long, 2023). If communication was removed or changed, the stability of the psychological variable would be lost (to some extent) and vice versa.

Although the decision to consider other causal variables in the measurement of stability is primarily a theoretical one, another potentially problematic assumption of the quasi-simplex model is more technical. The lag-1 assumption means that, in a three-wave model, T_3 is assumed to be

$$T_3 = \beta_{32}T_2 + u_3$$

Or expressed differently,

$$T_3 = \beta_{32}T_2 + \beta_{31}T_1 + u_3, \qquad \beta_{31} = 0$$

This formalizes the meaning of the lag-1 assumption: There is no added effect of T_1 on T_3 besides T_1's impact on T_2. When the true value of $\beta_{31} \neq 0$, it may not be obvious what impact this has on the meaning or accuracy of stability measurement. First, it may mean that the estimate of β_{32} is incorrect and that wave-to-wave stability is wrongly estimated. More importantly, as shown by Egan (2020), estimates of reliability are biased downward if the lag-1 assumption is violated. As shown before, downward bias in reliability means upward bias in stability since true variation is mistaken for random error. When the time period between measures is short, it is more likely this assumption is violated since whatever process that leads to autocorrelation between measurements may not take place on such a short timescale. For instance, if there are daily measures but the behavior occurs on a regular, 1-week basis, the 6th lag is the only one with a non-zero effect for the day the behavior occurs. Egan (2020) shows evidence the lag-1 assumption is violated for several social identity measures in a 3-wave panel in which measures are spaced 2 years apart, so it is not as if fairly long time lags are perfect protection against violations of this assumption. Properties of X_t likely to cause a downward bias in reliability estimates include relatively high wave-to-wave correlations (i.e., $\beta_{t,t-1}$) and low between-subject variance at wave 1 (Egan, 2020).

Why would applied researchers continue to use methods that can be biased under such common conditions? Besides the added analytic and conceptual complexity associated with relaxing the lag-1 assumption (and others already mentioned), three panel waves of a single indicator are insufficient to identify a structural equation model that includes the β_{31} parameter in addition to reliability estimates. An approximate test of the lag-1 assumption is possible however, via estimating a regression equation of the form

$$X_3 = \hat{\beta}_{32}X_2 + \hat{\beta}_{31}X_1 + e_3$$

A value of $\hat{\beta}_{31}$ statistically distinct from zero suggests the lag-1 assumption has been violated. Unfortunately, it is not possible to both model the lag-2 effect and correct for measurement error. For the uses of the simplex model that I am most concerned with, however, the purpose is not to correct for attenuation bias but to simply describe the stability of a variable. In this case, it is worthwhile to — at minimum — check this assumption as one way to judge the trustworthiness of reliability and stability estimates. Conceptually, it is more difficult to reduce stability to a single number once the lag-1 assumption is relaxed. As is always true in regression modeling, adding another predictor increases the explained variance in X_3. This means when the lag-1 assumption is violated, there is *more* variation in X_3 and by assumption T_3 explained by past values. This might seem to be an argument that conventional stability estimates when the lag-1 assumption is violated *under*estimate stability, then. More important, however, is that this increases the estimate of $Var(T)$ and decreases the estimate of $Var(e)$, the measurement error. The takeaway about violated lag-1 assumptions is that the true variation in the construct is underestimated. Using the regression procedure on the same NAES data from Dilliplane and colleagues (2013), I can demonstrate (full details omitted for brevity) that the lag-1 assumption is very clearly violated for all plausible operationalizations of media use and partisan media use — this casts doubt on the validity of the reliability and stability estimates commonly produced about media exposure measures.

There are other issues that have ramifications for reliability estimation that I will not describe in such detail. Alwin (2007), for example, shows that for measures with fewer than about 16 response options, the use of Pearson correlations rather than polychoric correlations biased reliability estimates downward by about .1, which is far from trivial. The solution for calculating stability in the face of this problem is not so clear, however, because polychoric correlations cannot just be plugged into the Heise (1969) equations for calculating stability (Jagodzinski & Kühnel, 1987; Scharkow, 2019). The assumption mentioned only in passing earlier about a lack of serial correlation in the errors may be untenable in many applied settings as well. The effect of serial correlation is more difficult to predict in part because it is an additional source of variance that is neither random error e_t nor true score variance u_t, but it stands to reason that unmodeled positive serial error correlations likely inflate estimates of stability since these errors are probably picked up by the $\beta_{t,t-1}$ terms in the simplex model. J. A.

Wiley and M. G. Wiley (1974) propose an alternate model that models error correlations with different identifying restrictions, a model which Achen (1983) felt called into question his earlier claims of very high stability of attitudes (Achen, 1975). This model has not caught on to the same extent because it is suffers from a great deal of sampling error unless there are at least 4, and preferably 5 or more, measurement occasions (Palmquist & Green, 1992). Others still have discarded this general framework for longitudinal data in favor of approaches related to intraclass correlation focused on partitioning variation (Bland & Altman, 1996; Laenen et al., 2009). All these models also assume linear relations, which is hardly an unusual assumption in applied research, but doing so is bound to conflate cyclic within-person processes like those expected in a negative feedback system with random measurement error. I do not explore this problem in detail here because of its complexity and due to the high demands in terms of data to empirically differentiate cyclic longitudinal change from random error.

To review, there are a number of reasons to suspect estimates of stability using the quasi-simplex model may be biased upward. Although I have not proposed any simple alternative, because there does not seem to be one, I merely hope to increase awareness of the considerable uncertainty involved in estimating measurement error-corrected stability and stimulate discussion on alternative approaches. Going forward, readers may wish to supplement the quasi-simplex approach with additional methods, such as modeling lag-2 effects in regression models or using polychoric correlations, to gain a more comprehensive understanding of construct stability. Researchers should also consider the conceptual implications of their stability estimates and whether they align with the specific research questions being addressed.

References

Achen, C. H. (1975). Mass political attitudes and the survey response. *The American Political Science Review*, *69*(4), 1218–1231. https://doi.org/10.2307/1955282

Achen, C. H. (1983). Toward theories of data: The state of political methodology. In A. W. Finifter (Ed.), *Political science: The state of the discipline* (pp. 69–94). American Political Science Association. http://archive.org/details/politicalscience00hodo

Achen, C. H. (2000). *Why lagged dependent variables can suppress explanatory power of other independent variables*. Annual Meeting of the American Political Science Association, Los Angeles, CA.

Allison, P. D. (2009). *Fixed effects regression models*. SAGE Publications. https://doi.org/10.4135/9781412993869.d33

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*. Wiley-Interscience.

Alwin, D. F., & Krosnick, J. A. (1991). Aging, cohorts, and the stability of sociopolitical orientations over the life span. *American Journal of Sociology*, *97*(1), 169–195. https://doi.org/10.1086/229744

Bland, J. M., & Altman, D. G. (1996). Statistics Notes: Measurement error and correlation coefficients. *BMJ*, *313*(7048), 41–42. https://doi.org/10.1136/bmj.313.7048.41

Clark, T. S., & Linzer, D. A. (2015). Should I use fixed or random effects? *Political Science Research and Methods*, *3*(2), 399–408. https://doi.org/10.1017/psrm.2014.32

Converse, P. E. (1962). Information flow and the stability of partisan attitudes. *Public Opinion Quarterly*, *26*(4), 578–599. https://doi.org/10.1086/267129

Converse, P. E. (1964). The nature of belief systems in mass publics. In D. E. Apter (Ed.), *Ideology and Discontent* (pp. 206–261). Free Press. http://ftp.voteview.com/The_Nature_of_Belief_Systems_in_Mass_Publics_Converse_1964.pdf

de Vreese, C. H., Boukes, M., Schuck, A., Vliegenthart, R., Bos, L., & Lelkes, Y. (2017). Linking survey and media content data: Opportunities, considerations, and pitfalls. *Communication Methods and Measures*, *11*(4), 221–244. https://doi.org/10.1080/19312458.2017.1380175

Dilliplane, S. (2014). Activation, conversion, or reinforcement? The impact of partisan news exposure on vote choice. *American Journal of Political Science*, *58*(1), 79–94. https://doi.org/10.1111/ajps.12046

Dilliplane, S., Goldman, S. K., & Mutz, D. C. (2013). Televised exposure to politics: New measures for a fragmented media environment. *American Journal of Political Science, 57*(1), 236–248. https://doi.org/10.1111/j.1540-5907.2012.00600.x

Egan, P. J. (2020). Identity as dependent variable: How Americans shift their identities to align with their politics. *American Journal of Political Science*, *64*(3), 699–716. https://doi.org/10.1111/ajps.12496

Eveland, W. P., Jr., Hutchens, M. J., & Shen, F. (2009). Exposure, attention, or "use" of news? Assessing aspects of the reliability and validity of a central concept in political communication research. *Communication Methods and Measures*, *3*(4), 223–244. https://doi.org/10.1080/19312450903378925

Garrett, R. K., Gvirsman, S. D., Johnson, B. K., Tsfati, Y., Neo, R., & Dal, A. (2014). Implications of pro- and counterattitudinal information exposure for affective polarization. *Human Communication Research*, *40*, 309–332. https://doi.org/10.1111/hcre.12028

Gervais, B. T. (2014). Following the news? Reception of uncivil partisan media and the use of incivility in political expression. *Political Communication*, *31*, 564–583. https://doi.org/10.1080/10584609.2013.852640

Green, D. P., & Palmquist, B. (1990). Of artifacts and partisan instability. *American Journal of Political Science*, *34*(3), 872–902. JSTOR. https://doi.org/10.2307/2111402

Guess, A. M. (2015). Measure for measure: An experimental test of online political media exposure. *Political Analysis*, *23*(1), 59–75. https://doi.org/10.1093/pan/mpu010

Heise, D. R. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review*, *34*(1), 93–101. https://doi.org/10.2307/2092790

Heywood, H. B. (1931). On finite sequences of real numbers. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *134*(824), 486–501. https://doi.org/10.1098/rspa.1931.0209

Jagodzinski, W., & Kühnel, S. M. (1987). Estimation of reliability and stability in single-indicator multiple-wave models. *Sociological Methods & Research*, *15*(3), 219–258. https://doi.org/10.1177/0049124187015003003

Jürgens, P., Stark, B., & Magin, M. (2019). Two half-truths make a whole? On bias in self-reports and tracking data. *Social Science Computer Review*, 0894439319831643. https://doi.org/10.1177/0894439319831643

Kruikemeier, S., & Shehata, A. (2017). News media use and political engagement among adolescents: An analysis of virtuous circles using panel data. *Political Communication*, *34*(2), 221–242. https://doi.org/10.1080/10584609.2016.1174760

LaCour, M. J., & Vavreck, L. (2014). Improving media measurement: Evidence from the field. *Political Communication*, *31*(3), 408–420. https://doi.org/10.1080/10584609.2014.921258

Laenen, A., Alonso, A., Molenberghs, G., & Vangeneugden, T. (2009). Reliability of a longitudinal sequence of scale ratings. *Psychometrika*, *74*(1), 49–64. https://doi.org/10.1007/s11336-008-9079-7

Lee, C., Hornik, R., & Hennessy, M. (2008). The reliability and stability of general media exposure measures. *Communication Methods and Measures*, *2*(1–2), 6–22. https://doi.org/10.1080/19312450802063024

Long, J. A. (2023). Stability as an outcome in communication research. *International Journal of Communication*, *17*, 5954–5971.

Long, J. A., Eveland, W. P., Jr., & Slater, M. D. (2019). Partisan media selectivity and partisan identity threat: The role of social and geographic context. *Mass Communication and Society*, *22*(2), 145–170. https://doi.org/10.1080/15205436.2018.1504302

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley Pub. Co.

Menon, G. (1993). The effects of accessibility of information in memory on judgments of behavioral frequencies. *Journal of Consumer Research*, *20*(3), 431–440. https://doi.org/10.1086/209359

Menon, G. (1997). Are the parts better than the whole? The effects of decompositional questions on judgments of frequent behaviors. *Journal of Marketing Research*, *34*(3), 335–346. https://doi.org/10.2307/3151896

Moeller, J., Shehata, A., & Kruikemeier, S. (2018). Internet use and political interest: Growth curves, reinforcing spirals, and causal effects during adolescence. *Journal of Communication*, *68*(6), 1052–1078. https://doi.org/10.1093/joc/jqy062

Palmquist, B., & Green, D. P. (1992). Estimation of models with correlated measurement errors from panel data. *Sociological Methodology*, *22*, 119. https://doi.org/10.2307/270994

Prior, M. (2013). The challenge of measuring media exposure: Reply to Dilliplane, Goldman, and Mutz. *Political Communication*, *30*(4), 620–634. https://doi.org/10.1080/10584609.2013.819539

Scharkow, M. (2019). The reliability and temporal stability of self-reported media exposure: A meta-analysis. *Communication Methods and Measures*, *13*(3), 198–211. https://doi.org/10.1080/19312458.2019.1594742

Schwarz, N., Hippler, H.-J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, *49*(3), 388–395. https://doi.org/10.1086/268936

Shehata, A., Ekström, M., & Olsson, T. (2016). Developing self-actualizing and dutiful citizens: Testing the AC-DC model using panel data among adolescents. *Communication Research*, *43*(8), 1141–1169.

Wheaton, B., Muthén, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. *Sociological Methodology*, *8*, 84–136. https://doi.org/10.2307/270754

Wiley, D. E., & Wiley, J. A. (1970). The estimation of measurement error in panel data. *American Sociological Review*, *35*(1), 112–117. https://doi.org/10.2307/2093858

Wiley, J. A., & Wiley, M. G. (1974). A note on correlated errors in repeated measurements. *Sociological Methods & Research*, *3*(2), 172–188. https://doi.org/10.1177/004912417400300202

Wojcieszak, M., Azrout, R., Boomgaarden, H., Alencar, A. P., & Sheets, P. (2017). Integrating Muslim Immigrant Minorities: The Effects of Narrative and Statistical Messages. *Communication Research*, *44*(4), 582–607. https://doi.org/10.1177/0093650215600490