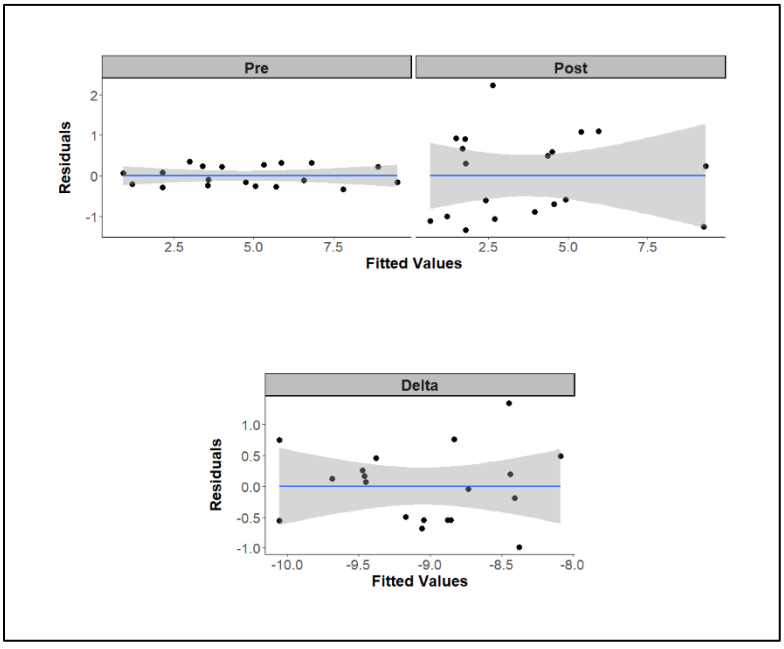
Supplementary Material: Issues to Consider when Comparing the Accelerometer-Based Intake-Balance Method against the NIDDK Body Weight Planner

In the main text of this paper, we provided proof-of-concept for an accelerometer-based intake-balance method of assessing energy intake (EI) in the context of a time restricted eating (TRE) intervention. As part of our demonstration process, we compared EI estimates from the accelerometer-based method against estimates from the Body Weight Planner of Hall et al.[1], which is hosted by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK; see niddk.nih.gov/bwp). Because both methods used some of the same information when predicting EI, our analysis cannot be considered a true validation. In particular, the shared information could cause the results to reflect artificial agreement rather than true validity. On the other hand, many other factors were also at play in the analysis, which could attenuate the impact of shared information. These contrasting possibilities warrant further commentary from statistical, methodological, and empirical perspectives. The purpose of this supplement is to explore each of those areas, with particular attention to implications for interpreting our results.

1. Statistical Integrity of the Analyses

The first potential issue to explore relates to the lack of independence between the two methods, resulting from their use of partially overlapping information. In particular, it is important to consider whether the shared information led to violation of any statistical assumptions in our tests. (Notably, this question of statistical integrity is separate from the question of empirical integrity, since a test can be statistically valid yet empirically nuanced. We will address empirical integrity in a later section.)

Our statistical tests were T-tests, equivalence testing, and regression analyses, which primarily require independence of observations rather than independence of measures. In our tests, each participant represented an independent observation, and thus no assumptions of independence were violated. Furthermore, **the figure at right shows a scatterplot of model residuals (Y-axis) versus fitted values (X-axis) for each timepoint, when regressing estimates from the NIDDK method against estimates from the accelerometer-based method**. No discernible patterns are visible in any of the panels, and the blue lines (best fit from simple linear regression, with standard error shading) all have slope and intercept very close to zero. This

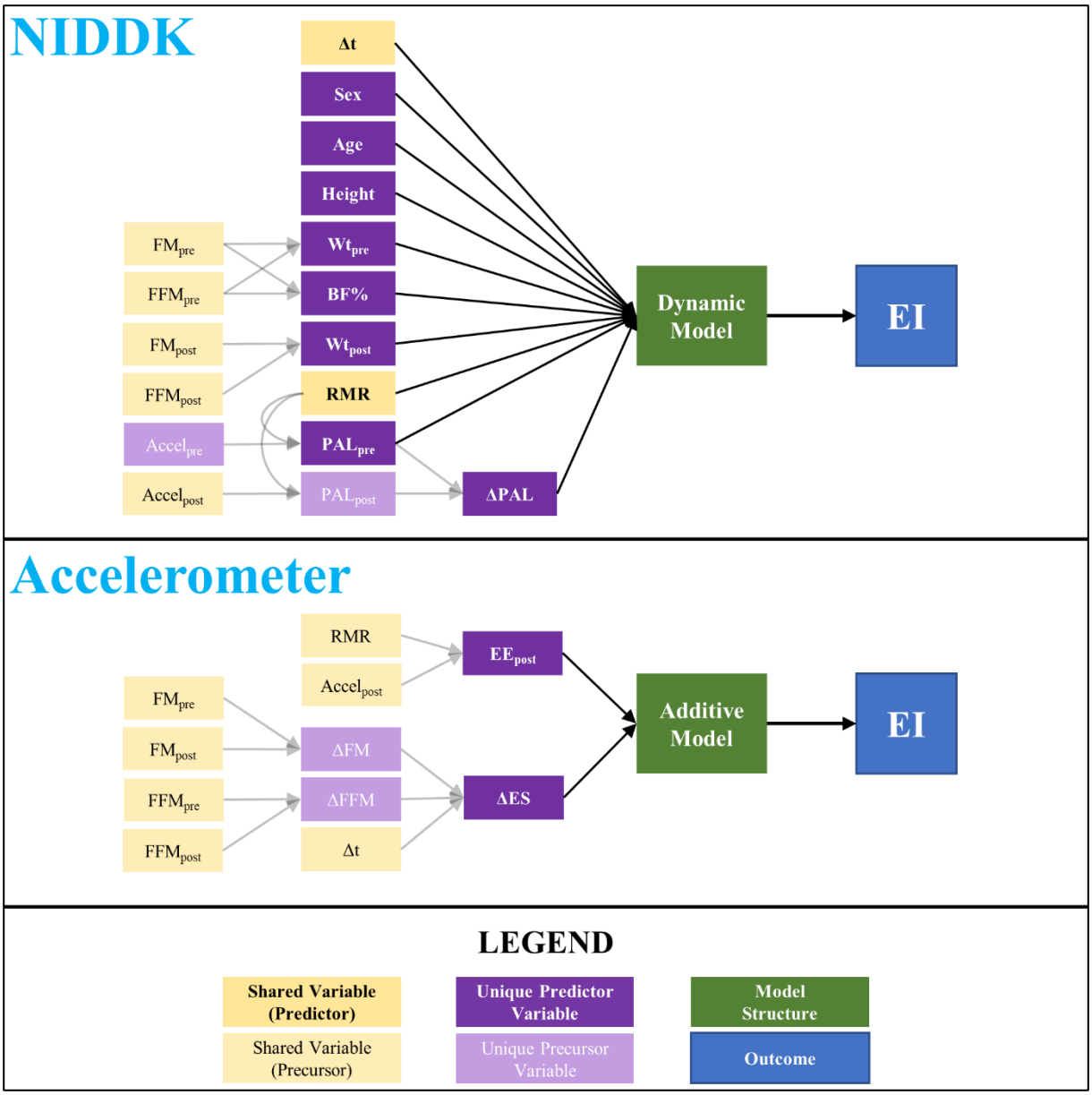


[1]Hall et al. (2011). Quantification of the effect of energy imbalance on bodyweight. *The Lancet*, 378(9793), 826-837. PMID: 21872751

underscores the minimal impact of shared information on the statistical integrity of our tests and models. In the following section, we will explore why this is, based on a comparison of how the accelerometer-based and NIDDK methods are structured and implemented.

2. Methodological Comparison of the Accelerometer-Based and NIDDK Methods

To understand how the two methods are related, it is important to consider their similarities and differences. These include not only the specific variables and interactions they each rely upon (some overlapping and others not), but also the specific transformations and weightings that are used, along with the overall modeling structures. These features are best compared schematically using the below figure.



Abbreviations: FM (fat mass), FFM (fat-free mass), Accel (accelerometer), Wt (bodyweight), BF % (bodyfat percentage), RMR (resting metabolic rate), PAL (physical activity level), EI (energy intake).

The schematics reveal considerable differences in how the two methods operate, which likely affects the impact of overlapping information. For example, the NIDDK method involves many more predictors than the accelerometer-based method, requiring greater nuance when interpreting the influence of any one variable in isolation. That is, a change in one variable for either method would be interpreted in terms of its impact "when controlling for the other variables in the model", and the extent of required control would be much greater for the NIDDK method than the accelerometer-based method. Another key observation is that most overlapping variables are precursors to the actual predictor variables used in each model. In other words, the predictor variables are mostly distinct. This may be especially powerful for uncoupling the two methods when combined with the differences in modeling structure (i.e., adaptive modeling for the NIDDK method versus additive modeling for the accelerometer-based model). Overall, these types of differences between the two methods likely played a key role in promoting the statistical integrity that was documented in the prior section (particularly the plots of residuals versus fitted values). However, questions remain surrounding the empirical value of comparing two methods that share some underlying information, as well as how to interpret comparisons of two such methods. We address this topic in the following section.

3. Empirical Integrity of the Analyses

As noted earlier, an analysis can be statistically valid yet empirically nuanced. In this case, the main concern (regardless of the statistical soundness of the tests and the methodological distinctions between the methods) is that the partially shared information could create an exaggerated picture of agreement, which could be misconstrued to reflect validity of the accelerometer-based method. While it is not possible to directly address the presence or magnitude of such an effect, there are nevertheless some critical points that can help to clarify the meaning and empirical value of the analysis.

The first thing to note is that the accelerometer-versus-NIDDK analyses cannot be construed as true validation of the accelerometer-based method, for two reasons: First, a true validation would require testing the accelerometer-based method against a fully independent one, which disqualifies the NIDDK method; and second, even if the NIDDK method were fully independent of the accelerometer-based method, it is not a gold standard measure, meaning tests of agreement would only reflect convergent validity, not criterion validity.

Nevertheless, it is also important to note that our analyses focused on magnitude of bias rather than shared variance and correlation. This is important because two methods can exhibit considerable differences (i.e., bias) even in the presence of shared variance[2]. Thus, although the overlapping information between the accelerometer-based and NIDDK methods would presumably create some shared variance, it would not necessarily cause low bias (our primary interest). Combined with the evidence presented in prior sections, this may suggest the shared information had neither a strong nor a systematic influence on the level of agreement we observed. However, we emphasize again that the full quantitative impact of shared information cannot be directly addressed.

[2]Bland & Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476), 307-310.

Lastly, it is noteworthy to compare estimates from both methods against prior literature. Although this does not directly inform how the agreement metrics are interpreted from our comparative analysis, it does provide indirect evidence of whether the predicted values are plausible, thereby providing greater context to the analysis. We noted in the Discussion section of the main text that prior studies have shown TRE interventions to result in EI reductions of 8%-20%, and that those reductions are comparable with our findings for the TRE group when using the accelerometer-based method (9.9% ± 6.4%) and the NIDDK method (12.3% ± 2.9%). Thus, the methods appear to have reasonable sensitivity to change. Additionally, the raw EI estimates from both methods (see Table 2 in the main text) appear to be biologically and behaviorally plausible in comparison with other studies[3], particularly when accounting for the underestimations that are associated with commonly-used self-report assessment tools[4]. Therefore, our results seem to indicate plausible estimates, although further validation is needed.

4. Conclusion

While the results of our analysis cannot be interpreted as validating the accelerometer-based method, they provide proof-of-concept that helps to build the case for ongoing work. By comparing accelerometer-based estimates against the NIDDK method, we were able to provide demonstration and context for the new method and show the general plausibility of its estimates. Despite the clear need for more research that directly validates and refines the accelerometer-based method, such ongoing work will greatly benefit from the foundation laid in this study.

[3]Wakimoto & Block (2001). Dietary intake, dietary patterns, and changes with age: an epidemiological perspective. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 56(suppl_2), 65-80.
[4]McClung et al. (2018). Dietary intake and physical activity assessment: current tools, techniques, and technologies for use in adult populations. *American Journal of Preventive Medicine*, 55(4), e93-e104.