Received: 28 November 2021 | Revised: 29 November 2023 | Accepted: 11 December 2023

DOI: 10.1002/job.2771

RESEARCH ARTICLE

WILEY Journal of Organizational Behavior

Time and change: A meta-analysis of temporal decisions in longitudinal studies

Helen Hailin Zhao[1] | Abbie J. Shipp[2] | Kameron Carter[3] | Erik Gonzalez-Mulé[4] | Erica Xu[5]

[1]Faculty of Business and Economics, The University of Hong Kong, Pok Fu Lam, Hong Kong

[2]Neeley School of Business, Texas Christian University, Fort Worth, Texas, USA

[3]Department of Management, Old Dominion University, Norfolk, Virginia, USA

[4]Kelley School of Business, Indiana University, Bloomington, Indiana, USA

[5]School of Business, Hong Kong Baptist University, Kowloon Tong, Hong Kong

Correspondence
Helen Hailin Zhao, Faculty of Business and Economics, The University of Hong Kong, Pok Fu Lam, Hong Kong.
Email: hhzhao@hku.hk

Funding information
This paper is partially supported by the HKSAR General Research Fund (17500821) and the National Science Foundation of China (72002184). Erik Gonzalez-Mulé's time spent on this project was partially supported by the Randall L. Tobias Chair in Leadership from Indiana University's Kelley School of Business.

Summary

Longitudinal research has grown in popularity in the field of management and organizations. However, the literature has neglected to consider the important ways in which researchers' temporal decisions can influence observed change in longitudinal studies. Researchers must make a set of temporal decisions to capture change, such as the temporal precision of the hypothesized form of change, the selection of a sample that is expected to exhibit the change, the choice of variables to be measured repeatedly, the frequency of measurements, and the time interval between measurements. However, these decisions typically are based on "educated guesses," which makes their effects on the observed change unclear. In this paper, we develop a conceptual framework to explain how temporal decisions influence observed change and validate it by meta-analyzing longitudinal studies ($k = 268$). Specifically, we found that observed change is affected by hypotheses (i.e., temporal precision), the sample (i.e., presence of a change trigger), variables (i.e., variable type and rating source), and measurement occasions (i.e., frequency and time interval). These findings offer insights into the importance of making informed temporal decisions. The implications of our findings are broad and applicable across research streams and theoretical traditions.

KEYWORDS

meta-analysis, panel and repeated measure designs, research design, temporal decisions, time

1 | INTRODUCTION

Time serves as a frame of reference that enables people to see and explain changes (Epstein, 1979; McGrath, 1988; Shipp & Fried, 2014). In longitudinal studies, time is not a substantive construct itself but rather a temporal marker of other substantive constructs at different measurement points. The repeated measurements of a construct collected from the same respondents over time form the basis of longitudinal research, which allows researchers to observe the process of change.

All longitudinal studies, regardless of their research questions and analytical approaches, can be reduced to a relationship between time and change. Each longitudinally studied variable that is measured t times can be mathematically represented by a total of $[t \times (t-1)]/2$ difference scores (i.e., Cohen's d), which captures the observed change in the variable between any two given time points. This relationship can also be represented graphically, with t measurement points on the X-axis separated by $t-1$ time intervals. The level of the variable at each point in time corresponds to these t time points on the Y-axis. Drawing these points together yields an observed change trajectory that reflects how the variable changes over time. However, the shape of this change trajectory could be influenced by time-related variables on the X-axis, such as the frequency of repeated measures or the time intervals among the measurement points. In fact, it has been argued that poor temporal designs can lead to inaccurate conclusions (Ployhart & Vandenberg, 2010). For example, a mismatch between the measured change and the true change can lead to *temporal* Type I and Type II errors, analogous to the well-known Type I and Type II errors in statistics. A temporal Type I error captures a change that does not exist, whereas a temporal Type II error fails to capture a change that does exist. To reduce these potential errors, researchers must take extra care to consider time and change when developing theories and designs for longitudinal studies. Temporal decisions are critical to the conclusions we make in longitudinal studies.

Interestingly, although there has been an exponential growth in the number of longitudinal studies in the field of management and organization (Cortina et al., 2017), this type of research has been criticized because temporal decisions continue to be based on "intuition, chance, convenience, or tradition" (Mitchell & James, 2001, p. 533). As Pitariu and Ployhart (2010) observed, there is a lack of temporal precision in the hypothesis development within longitudinal studies. Even worse, Ployhart and Vandenberg (2010) noted that justifications for, or even descriptions of, temporal designs are often left out of longitudinal studies. Despite repeated calls for investigation, the challenge of temporal decisions in longitudinal research studies persists (Taris & Kompier, 2014).

We contend that the field of management needs a base of cumulative knowledge about theoretical and methodological choices covering time and change in longitudinal studies. Of course, given that each primary study comes from its own domain, the research design should be appropriate to the chosen research stream (e.g., leadership or teams) and theoretical perspective (e.g., attribution theory or social exchange theory). However, as we will demonstrate, research streams and theoretical perspectives are not enough to guide temporal decisions. The conventional wisdom in our field is that research design should be based on "theory" and "context," but theory and context often do not specify how temporal decisions should be made. Therefore, to gain a deeper understanding of the temporal aspect, we need a conceptual framework that explicitly addresses the theoretical development and methodological choices around time. The acquisition of this knowledge; however, is likely beyond the scope of a single field study, because each study only makes one set of temporal decisions and uncovers one set of observed change trajectories. Thus, meta-analysis becomes advantageous for examining the consequences of temporal decisions across the landscape of primary studies (Rosenthal & DiMatteo, 2001).

In this paper, we meta-analyze how temporal decisions influence observed change (e.g., Cohen's *d* of the change between two time points), using specific temporal decisions made by each primary study to predict the magnitude of the effect size (Borenstein et al., 2009). For example, if time intervals are long in some studies but short in others, a meta-analysis can show how the effect size of observed change differs based on this choice of time interval. By doing so, meta-analysis can create knowledge across studies that will help researchers make more informed temporal decisions in future longitudinal studies.

In our conceptual framework, we highlight important theoretical and methodological decisions about time in hypotheses, samples, variables, and measurement occasions. Our meta-analytic results reveal relationships between each of these categories of temporal decisions and the observed amount of change across time. By creating a set of theoretically driven and evidence-based principles, we contribute to longitudinal research in two important ways. First, our conceptual framework offers theoretical recommendations for developing research hypotheses with temporal precision. Second, it offers specific methodological recommendations on samples, variables, and measurement occasions to make more informed design decisions in future research that test the relationship of interest appropriately.

2 | A CONCEPTUAL FRAMEWORK OF TEMPORAL DECISIONS AND OBSERVED CHANGE

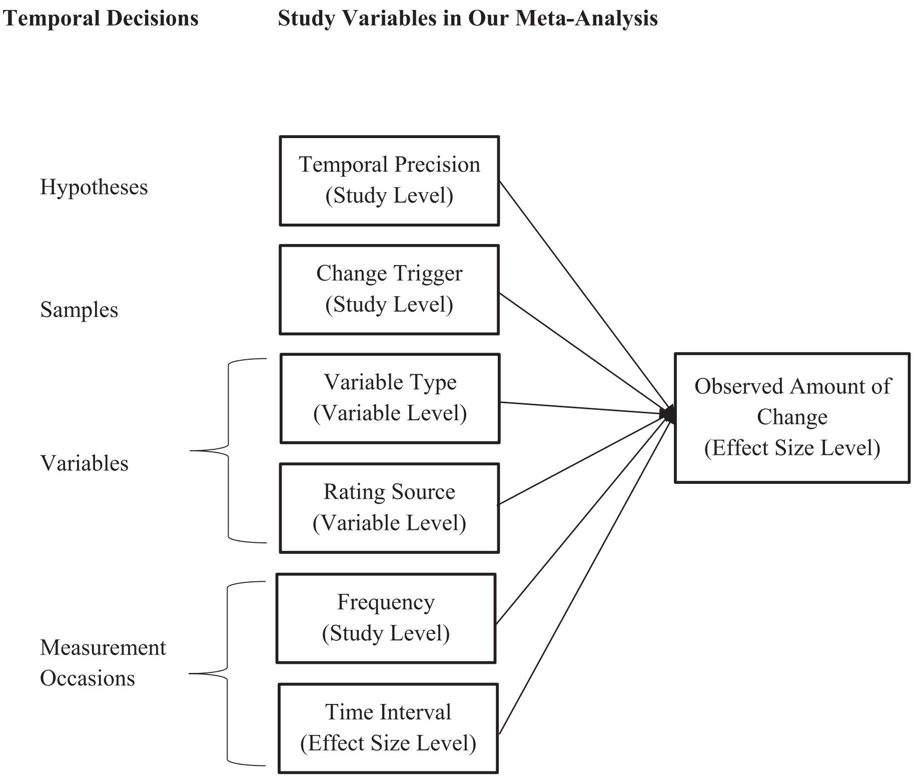
As shown in Figure 1, we build on prior conceptual work (Ployhart & Vandenberg, 2010; Zapf et al., 1996) to identify four categories related to temporal theory development and research design: hypotheses, samples, variables, and measurement occasions. These four categories form the basis of our framework for how temporal decisions influence observed change. Specifically, when designing a longitudinal study, researchers must develop *hypotheses* that precisely theorize the change over time, find an appropriate *sample* to capture the proposed change, choose *variables* of theoretical interest that are expected to change, and map *measurement occasions* onto a time scale to best measure the intended variables. Although not completely exhaustive, this framework covers the most important temporal decisions that can be found in longitudinal studies and quantified in a meta-analysis.

This framework should inform important temporal decisions for longitudinal researchers, including how the amount of change detected might be influenced by (1) the temporal precision in hypothesis development (e.g., the degree to which a change-related hypothesis has specified its timing, duration, and shape); (2) the selection of a sample with a change trigger (e.g., sampling near an organizational change versus a convenience sample without reason to change); (3) the choice of variable type (e.g., performance or attitude) and rating source (e.g., objective measures or subjective ratings); as well as (4) the frequency of measurement within the study duration and the length of the time interval between measurement occasions. In the sections that follow, we review each of these decisions, developing a set of hypotheses that predict how such temporal decisions influence observed change. As a meta-analysis, our goal is to inform broad research questions and provide general guidelines that apply across studies rather than granular decisions unique to a primary study. We will return to this point in the discussion section.

2.1 | Hypotheses

The first temporal decision for longitudinal research is to theorize the expected form of change with some degree of *temporal precision*,

FIGURE 1 The conceptual framework of temporal decisions and observed change.


Temporal Decisions
Study Variables in Our Meta-Analysis
Hypotheses
Samples
Variables
Measurement Occasions
Temporal Precision (Study Level)
Change Trigger (Study Level)
Variable Type (Variable Level)
Rating Source (Variable Level)
Frequency (Study Level)
Time Interval (Effect Size Level)
Observed Amount of Change (Effect Size Level)

defined as the amount of information provided about the temporal aspect of the change-related hypotheses. The biggest strength of the longitudinal design is that it can demonstrate the temporality of the relationship (Voelkle & Oud, 2015). As such, specificity around time is critical in the hypothesis development of longitudinal studies (Mitchell & James, 2001). Pitariu and Ployhart (2010) provide a framework to evaluate the temporal precision of change-related hypotheses. They propose a continuum of strong to weak hypotheses with three criteria: timing,[1] duration, and shape. Timing denotes when a change occurs or when a relationship exists within the flow of time. Duration specifies how long a relationship is expected to last. Shape involves describing the functional form of the relationship over the course of investigation.

Although longitudinal designs have been increasingly employed, many of the primary studies simply state in their hypotheses that one variable is positively or negatively related to another. Pitariu and Ployhart (2010) argue that these hypotheses lack temporal precision, indicating that theoretical arguments leading to the hypotheses did not fully explore the temporal aspect of the relationship. Some of the hypotheses specify the timing of the relationship, such as "over time," "over the weekend," or "after an event" (e.g., Fritz et al., 2010; Kiburz et al., 2017). While the level of temporal precision is still low, it represents at least some improvement compared to hypotheses that do not specify timing, duration, and shape at all. In contrast, other studies have a moderate level of temporal precision because they theorize that the change of one variable triggers the change of another variable over time, aiming to understand why the change occurs and how it manifests itself over time (e.g., Toker & Biron, 2012). Finally, hypotheses with a high level of temporal precision cover timing, duration, and shape (Mitchell & James, 2001; Pitariu & Ployhart, 2010). An example is from Lorinkova and Bartol (2021), which hypothesizes that "Shared leadership will change over time following a curvilinear pattern, increasing during Phase 1 of project team development, peaking around the mid-point, and decreasing during Phase 2. This curvilinear pattern, approximating an inverted U-shape, will relate positively to team performance." A high level of temporal precision indicates that researchers have thoroughly theorized the temporal nature of the change to develop for why, how, and when a variable should change over time.

We argue that developing theories with greater temporal precision requires researchers to understand such temporal issues more deeply in their theoretical arguments (Mitchell & James, 2001). A higher level of temporal precision in one's hypotheses should then lead to more informed and appropriate decisions about temporal design, allowing researchers to capture more of the change they seek in their analyses. We thus hypothesize that in longitudinal studies focused on capturing change, a higher level of temporal precision in the research hypotheses should lead to a greater amount of observed change.

> **Hypothesis 1.** Observed change is greater in longitudinal studies that offer hypotheses that are stronger in temporal precision.

[1] Pitariu and Ployhart (2010) refer to this concept as "time"; however, we use the term "timing" to more specifically reflect their concept, particularly how its distinction from the more general term "time" that is typically used quite broadly.

2.2 | Samples

Having specified the form of change, it is important to choose an appropriate sample to capture this theorized form. The conventional wisdom is that, whenever possible, researchers should choose a sample that is expected to exhibit the theorized form of change. A convenience sample that lacks relevance to questions of interest is clearly not advisable, and researchers should at least provide evidence for why the chosen sample is appropriate to test the change-related hypotheses (Ployhart & Vandenberg, 2010). In most cases, this means specifying some sort of event to provide a theoretical rationale for why change is expected in a particular study. However, despite being integral to detecting observed change in a focal construct, such triggers for change have not received sufficient scholarly attention. Convenience sampling remains common and many longitudinal studies even skip discussions of why change should be expected in a particular sample (Ployhart & Vandenberg, 2010; Rousseau & Fried, 2001).

We conceptualize that an important construct underlying this decision is the *change trigger*, defined as an event that takes place among research participants or in the research site that can potentially cause changes in the substantive construct of interest. For instance, to test a theory of retirees' adjustment processes, Wang (2007) sampled senior employees near retirement (i.e., the change trigger), tracking them before, during, and after leaving the workforce. Given that retirement is a meaningful event that triggers a change, any study that surveyed a randomly chosen group of people (e.g., those earlier in their careers) obviously would be less likely to uncover different patterns in employees' adaptation processes. Beyond this example, many other types of change triggers exist in management research, such as changing jobs (van der Werff & Buckley, 2017), transitioning to self-directed teams (Douglas & Gardner, 2004), or attending a training session (Strauss & Parker, 2018).

We propose that using samples with an explicit change trigger is critical to understanding a hypothesized change because it specifies a contextual driver of change at a point in time. In contrast, when there is no event to instigate a change, it is difficult to justify a hypothesized change form. According to inertial effects (Hannan & Freeman, 1984), when there is no strong reason for people to change, they tend to remain unchanged or do nothing. Research on resistance to change also documents people's tendency to maintain the status quo out of perceptions of threat, fear of uncertainty, or difficulty adapting (Oreg, 2003). As a result, organizational samples that do not specify a change trigger, such as convenience samples, may inadvertently capture stability for some individuals but change for others, reducing the overall amount of change detected over time (if any change is detected at all). Therefore, we contend that the observed amount of change in a longitudinal study should be greater if the study specifies an explicit change trigger, which initiates some form of change.

Hypothesis 2. Observed change is greater in studies with a specified change trigger.

2.3 | Variables

The next temporal decision for longitudinal research relates to the choice of variables one expects to change. For an empirical study, the choice of the variable(s) of interest is typically driven by a specific theory within the context of its own research area. In our framework, we view these variables from a perspective of how they are conceptualized temporally. By their very nature, some variables are more likely to change than others; however, we are not aware of any theoretical or empirical research that tests this idea across different variable types. As an initial investigation, we identify two theoretical concepts related to the choice of a variable: *variable type*—the nature of the variable, such as attitude or behavior; and *rating source*—the party who provides the rating of the variable, such as the self or a supervisor.

It is important to note that, for the purposes of our theorizing below, we refer to theory and findings from various levels (e.g., individual, team, and organization). However, given that we employ meta-analytic methods and are interested in reporting trends across the literature, we are unable to examine more fine-grained patterns (e.g., oscillating patterns) or make conclusions about units at a lower level than the aggregated data we obtained from our study (e.g., individuals); thus, our hypotheses are framed around the level we examine in the meta-analysis, which we extracted from study-level data.

2.3.1 | Variable type

It makes intuitive sense that certain variable types are more likely to change than others, but almost no studies have directly compared how the magnitude of change differs across *variable types*. Common variable types in the field of management research include personal attributes (e.g., personality or cognitive abilities), performance (e.g., task performance or extra-role performance), non-performance behaviors (e.g., job search behaviors or coping behaviors), attitudes (e.g., job satisfaction or organizational commitment), and emotions (e.g., positive and negative affect; for a similar categorization of variable types, see Bosco et al., 2015 and O'Boyle et al., 2019). Among these variable types, personal attributes are presumed to be governed by temperament or genetic factors, making them less likely to change. For example, Judge et al. (1999) found that personality is so stable that it can be used to predict career success across one's life span. Similarly, Roberts, Walton, and Viechtbauer's (2006) meta-analysis also investigated personality across the life course, finding that although change can occur, personality shows substantial stability over time. Based on these findings, we conclude that personal attributes likely exhibit a relatively high degree of temporal consistency (Costa & Mccrae, 1999).

By comparison to the relative stability of personal attributes, other variable types such as performance, non-performance behaviors, attitudes, and emotions should be more susceptible to change. However, it is not yet clear how change over time may differ among these variable types. Each has its own well-established stream of research, making it challenging to draw definitive conclusions. Yet, based on theories and the findings of prior empirical work, we develop a set of predictions about the magnitude of change across variable types.

First, we propose that performance should exhibit less change than non-performance behaviors. This is because task performance is subject to many constraints including motivation, ability, effort, job demands, and even luck (Anderson & Butzin, 1974; Liao & Chuang, 2004; Lim & Tai, 2014), some of which are beyond an individual's control. To this point, Farrell and McDaniel (2001) and Ployhart and Hakel (1998) both showed that performance trends follow a learning curve; after passing a certain point in time, one's task performance tends to stabilize. In contrast, although non-performance behaviors also can be constrained (e.g., finding the right time to help a colleague), they should have fewer constraints than performance. That is, although individuals do not necessarily have full control about the level of productivity, they have more control about whether or not they want to help others. Further, such non-performance behaviors are less often evaluated against a specific, well-established standard, making them more likely to vary compared to task performance. For this reason, we predict that performance may exhibit less observed change than non-performance behaviors, because non-performance behaviors continue to fluctuate as situations unfold.

In addition, moving along the spectrum of variable types, we argue that non-performance behaviors exhibit less change than attitudes. Whereas non-performance behaviors are constrained by the situational factors in the specific behavioral context (e.g., the target, history, or norms of helping in a particular workgroup), attitudes are not as situationally constrained. They may fluctuate more rapidly as evaluative thoughts and judgments ebb and flow (Ajzen & Fishbein, 1977; Ehrlich, 1969). Thus, we expect that attitudes will change more frequently than non-performance behaviors.

Further, we reason that compared to attitudes, emotions change even faster. This is because emotion by definition is a discrete and short-lived phenomenon (Elfenbein, 2007). In other words, it is an involuntary, real-time affective reaction that has very few constraints. For example, whereas attitudes such as organizational commitment and job satisfaction are evaluative beliefs that take more time to form and evolve, the emotional response to disappointing news at work may change within a few seconds or minutes (Eagly & Chaiken, 1993). Thus, based on the existing research evidence, we expect emotions to fluctuate even more rapidly than attitudes, making them the variable type most likely to exhibit observed change.

In sum, we hypothesize that observed change in longitudinal studies will be lowest for variables that measure personal attributes, followed by performance, then non-performance behaviors, then attitudes, and finally emotions, which should exhibit the most amount of change.

> **Hypothesis 3.** Observed change varies across variable types, with an ascending amount of change expected for variables that measure personal attributes (least change), performance, non-performance behavior, attitudes, and emotions (most change).

2.3.2 | Rating source

The other important aspect to understanding a variable from a temporal perspective is its *rating source*. Prior research demonstrates that variance attributable to different rating sources is not simply measurement error but instead offers divergent yet meaningful perspectives about the subject being rated (Hoffman et al., 2010; Lance et al., 1992; Woehr et al., 2005).

One important distinction with respect to rating source is the difference between subjective ratings and objective measures of variables. Subjective ratings are those reported by oneself or others (e.g., supervisor, peer, or subordinate), whereas objective measures are impartial reports of indices or metrics typically retrieved from archival data. In the case of job performance ratings, Bommer et al. (1995) noted that objective and subjective measures cannot be used interchangeably because they correlate only at 0.32. Based on meta-analytic findings, Sturman et al. (2005) echoed this finding and further showed that subjective measures of performance changed less than objective measures. Given that subjective measures had higher correlations over time than did objective measures, this finding could indicate that subjective rating sources exhibit more stable response biases (e.g., halo or horn effects) compared to objective measures that may fluctuate more over time. Thus, consistent with prior research, we make an overall prediction that objective ratings should exhibit more change than subjective ratings.

In addition, we make further predictions about the degree of change within the different types of subjective ratings. We build our theoretical predictions on the degree to which a rating source of a variable makes stable inferences about the rated subject. First, it has been documented by psychometrics research that self-ratings have the fewest halo errors (Holzbach, 1978). This means that when raters rate themself, they do not rely on a preconceived global, overall judgment so that the rating fluctuates more across different dimensions and across different time points. However, halo errors are more substantial in all other variable rating sources (i.e., supervisor ratings, peer ratings, and subordinate ratings). As a result, we hypothesize that self-rated variables should exhibit more change than other subjective rating sources. Second, Viswesvaran, Schmidt, and Ones's (2005) meta-analysis found that supervisor ratings are less likely to show stable inferences about the rated subject than rating sources from other individuals such as subordinates and peers. This is perhaps because the supervisory role is evaluative in nature and has more formal and informal opportunities to evaluate the rated subject's variability over time. We thus hypothesize that supervisor rating should exhibit more change than subordinate and peer ratings. Third, although no empirical evidence compares peer and subordinate ratings, we speculate that peer ratings should exhibit more change. This is because peer ratings are positively correlated with supervisor ratings, which are generally subject to more change over time, whereas subordinate ratings are quite different from supervisor ratings (Harris & Schaubroeck, 1988; Mount et al., 1998). This could be because subordinates do not have as many opportunities to observe the ratee performance as compared to peers or supervisors. Using this logic, subordinate ratings are expected to exhibit the least amount of observed change.

Based on the aforementioned logic, we predict that change will be lower for subjectively measured variables than objectively measured variables, with the least change detected from ratings received from subordinates, followed by increasing degrees of change reported by peer ratings, supervisor ratings, and self-rating. Then, in contrast to these subjective ratings, objective ratings should demonstrate the most change. Thus, we predict:

> **Hypothesis 4.** Observed change varies across rating sources, following the ascending amount of change expected from subordinate rating (least change), peer rating, supervisor rating, self-rating, and objective measures (most change).

2.4 | Measurement occasions

Having chosen the hypotheses, sample, and variables, the final temporal decision faced by longitudinal researchers is the measurement occasion. From a lens of temporality, this decision involves mapping repeated measures of the same variable to a time scale. Researchers must decide how long the overall study will be, including frequency—how many measurements will be collected during the study, and time interval—how much time must transpire between two repeated measurements.

2.4.1 | Frequency

Two factors define the *frequency* of measurement occasions: (1) the total number of measurements across a study and (2) the total duration of the study. As such, frequency is a study-level variable. In designing a longitudinal study, researchers must determine the number of measurements over an appropriate span of time to model the hypothesized form of change. Ideally, researchers should measure as frequently as possible to reduce the likelihood of missing the "true" change. The true change form is typically unknown, but frequent measurement can allow the researchers to obtain a more accurate depiction of the phenomena they are studying. However, researchers often must make compromises (Ployhart & Vandenberg, 2010). For example, there may be practical constraints such as a research site that does not allow researchers to measure as frequently as they desire, or the researcher might not have enough resources (e.g., time, money, or staff) to support more frequent data collection. More critically, researchers may intentionally limit frequency because they believe frequent measurement might alter a study's conclusions. For instance, an abundance of measurement waves could increase participant fatigue and reduce data quality. Further, participant attrition over multiple waves could reduce the power of the study, and if the attrition is not random, the study's conclusions might be biased (Lance et al., 2000).

Because of these constraints, we propose that not all studies measure change as frequently as possible. Yet, we predict that studies with higher frequency are less likely to miss important changes during the study period, enabling researchers to capture a larger magnitude of change over time (Ployhart & Vandenberg, 2010). In other words, if a study has a higher frequency, it is more likely to capture the maximum and the minimum of the true change form, thus capturing a bigger magnitude of change. We therefore hypothesize:

> **Hypothesis 5.** The relationship between observed change and frequency is positive.

2.4.2 | Time interval

Time interval refers to the length of the time between two repeated measurement occasions of the construct of interest. In contrast to frequency, which is at the study level, time interval is captured at the effect size level because a single study could have varying time intervals among measurement occasions. To explain time intervals, Zaheer et al. (1999) identified several types of time scales, with the two most relevant being the existence interval and the recording interval. The existence interval is the period in which the true change manifests itself over time. The recording interval refers to decisions made by researchers about how to measure this period of change, based on their understanding of the existence interval. In a typical longitudinal study, researchers do not know the existence interval before the data collection and therefore do not know if the recording interval is consistent with the existence interval.

Our concept of time interval is conceptually analogous with the recording interval as discussed in Zaheer et al. (1999). Thus, time interval serves as a window of observation to the actual, unknown change phenomenon. The change trajectory observed largely depends on how narrow or wide the time interval is for a given effect. When the time interval is too large, the effect of the variable of interest could wear off, or other variables could enter and confound the relationship. In contrast, when the time interval is too short, the captured change may not represent a complete picture of the true change (Mitchell & James, 2001). We hypothesize that the relationship between time interval and observed change is likely curvilinear (i.e., inverted U-shape), such that some point exists at which the highest level of change can be captured. This is in contrast to the time interval being too long or too short, which would make the observed amount of change smaller. Thus, we hypothesize:

> **Hypothesis 6.** The relationship between observed change and time interval follows an inverted U-shaped, curvilinear pattern.

3 | METHOD

3.1 | Literature search, inclusion criteria, and data structure

Given our broad research scope, we limited our search to eight highly ranked journals, including (alphabetically) *Academy of Management Journal*, *Administrative Science Quarterly*, *Journal of Applied Psychology*, *Journal of Management*, *Journal of Organizational Behavior*, *Organizational Behavior and Human Decision Processes*, *Organization Science*, and *Personnel Psychology*. This is a commonly used journal list in large-scale meta-analyses (Aguinis et al., 2011; Gonzalez-Mulé & Aguinis, 2018; Judge et al., 2007; Yu et al., 2016) and is a particularly appropriate sampling list for our study as these journals publish high-quality exemplars of longitudinal research designs with a selective peer review process. We searched for article abstracts that contained the keywords "longitudinal" or "repeated measures" published any time through 2022 (including online first articles). Of the 7380 articles published in these eight journals, 1967 of them contained the search terms we used.

To be included in our sample, articles had to meet several criteria. First, they had to report true longitudinal research that measured the same variable over at least three time points, each separated by an interval of time (Ployhart & Vandenberg, 2010). This criterion excluded time-lagged studies (i.e., studies that measure one variable at one point in time and a different variable at another) and studies that only measure a variable twice, as well as qualitative studies, meta-analyses, and theoretical papers. Second, as we are primarily interested in work settings, we excluded samples that focused on biological change (e.g., observed change in pulse or hormone levels). Third, we only retained studies that provided means, standard deviations, or other statistics (e.g., correlation coefficient *r*, univariate *t*, Cohen's *d*) that could be used to estimate the observed change for the longitudinally measured variable at all time points. Most experience sampling method (ESM) studies were excluded under this criterion (interested readers can refer to McCormick et al., 2020 for a review of ESM research). Despite the fact that ESM is a longitudinal design with intensive repeated measures, the descriptive statistics for each time point are often not separately reported in the majority of ESM studies. Specifically, we excluded 45 (88%) ESM studies in our screening effort as they reported statistics at the variable level rather than at the time level, as is customary in ESM research (see Dimotakis et al., 2011, for an example). We were unable to estimate the effect size for these studies for the purposes of our meta-analysis. There are, however, a few ESM study exceptions that reported descriptive statistics for different time points and were therefore included in our meta-analysis, including Barclay and Kiefer (2019), Eatough et al. (2016), Frank et al. (2022), Gonzalez-Mulé and Yuan (2022), and Meier et al. (2016).

Using these criteria, we obtained 204 articles that included 268 samples. Our data structure is nested at three levels: the study level ($k = 268$), variable level ($k = 873$), and effect size level ($k = 4812$). The study level *n* indicates that there were 268 different studies obtained from the literature search. Many studies measured more than one variable repeatedly, giving a total of 873 variables nested within the 268 studies. Further, for each variable, there are multiple effect sizes between all pairs of measurement occasions of the variable. For example, if a longitudinal study has three time points, it has three effect sizes for each variable measured longitudinally (i.e., T1 to T2, T2 to T3, and T1 to T3). If a longitudinal study has four time points, it has six effect sizes (i.e., T1 to T2, T1 to T3, T1 to T4, T2 to T3, T2 to T4, and T3 to T4). Thus, there are 4812 effect sizes nested within the 873 variables. As an example to illustrate the structure of our data, consider Ambrose and Cropanzano (2003), a primary study coded in our meta-analysis that measured four variables (procedural justice, organizational commitment, job satisfaction, and turnover intentions) three times with time intervals of 8 months (T1 to T2), 12 months (T2 to T3), and 20 months (T1 to T3). For this *one* study (i.e., Level 3), we coded *four* different variables (i.e., Level 2), each of which has *three* effect sizes representing the amount of change between measurement occasions (i.e., Level 1). The reference list of the coded articles and the coding sheet that follows the Meta-analysis Reporting Standards can be found in Data S2.

3.2 | Variables used in the meta-analysis

3.2.1 | Effect size level

The *observed amount of change* and *time interval* are captured at the effect size level. Observed change refers to the amount of change across a pair of measurement points. Two coders independently coded the mean and standard deviation of each variable of interest at all possible time points. We also coded the time interval as the months between the same pair of repeated measures corresponding to each observed change. For example, if a study measured job performance three times, with 1 month separating the first two measurements (T1 and T2) and 3 months separating the next two measurements (T2 and T3), we coded 1 month as the time interval for the observed change between T1 and T2, 3 months for the time interval between T2 and T3, and 4 months for the time interval between T1 and T3. The coding of this variable is mostly straightforward except that a time point is sometimes unclearly described as "a few hours" or "right after"; to be consistent, we code such descriptions as the minimum value (i.e., 0.01 months) in our coding sheet. As a point of reference, 369 of the effect sizes included time intervals of 1 week or less; 476 were between 1 week and 1 month; 1478 were between one and 6 months; 564 were between 6 months and 1 year; 1164 were between 1 and 3 years; 415 were between 3 and 5 years; and 346 effect sizes had a time interval greater than 5 years.

3.2.2 | Variable level

Variable type and *rating source* are captured at the variable level. For variable type, we assigned codes according to whether the measured variable represented one of five categories: personal attributes ($k = 61$; e.g., proactive personality, Li et al., 2014); performance ($k = 99$; e.g., job performance, Day et al., 2004); non-performance behaviors ($k = 198$; e.g., job search behaviors, Kammeyer-Mueller et al., 2005); attitudes ($k = 406$; e.g., job satisfaction and organizational commitment, Vandenberghe et al., 2017); or emotions ($k = 52$; e.g., positive and negative affect, Fugate et al., 2002). Variables that could not be assigned to any of these five categories were coded as "others"; $k = 57$), which often represented rare categories with small k (e.g., base pay, Harris et al., 1998; team-level organizational tenure, Kuypers et al., 2018). Again, variable type was coded by the two coders independently, and the initial agreement rate was 83% and the Cohen's Kappa value was 0.66. Further, the same coding procedure was used to code rating source as objective data ($k = 102$), self-report ($k = 681$), supervisor-rated ($k = 29$), subordinate-rated ($k = 32$), or peer-rated ($k = 13$). When the rating source was unclear or was a combination of multiple rating sources, we coded it as "others" ($k = 16$). We achieved 96% initial agreement on this coding, and the initial Cohen's Kappa value was 0.92. All disagreements were resolved through discussion. In hypothesis testing, we use effect coding for the variable type and rating source as it will "anchor" these weights to the grand mean (Alkharusi, 2012), which avoids the arbitrariness of specifying the reference group in dummy coding.[2]

3.2.3 | Study level

Temporal precision, *change trigger*, and *frequency* are captured at the study level. First, we coded the temporal precision of the set of hypotheses in the study based on Pitariu and Ployhart's (2010) framework that offers three criteria for evaluating the temporal precision: timing, duration, and shape. We coded temporal precision as "0" if a study did not have any hypotheses that specified timing, duration, or shape. We coded a study as "1" if it specified timing and as "2" if it specified timing and duration or timing and shape. If all three criteria were included, the coding was "3." If a longitudinal study did not develop any hypotheses, it was excluded (15.67%; $n = 42$). We conducted supplementary analyses and found that some of the primary studies without research hypotheses exhibit a high level of temporal precision in their theoretical arguments, while others do not, resulting in a wide range of observed changes. We speculate that there may be various reasons for not developing research hypotheses, and this may not necessarily be related to the level of temporal precision. In addition, since the theoretical rationale of our prediction does not apply to longitudinal studies that were designed to show stability (2.61%; $n = 7$), we only included longitudinal studies that were designed to observe change.[3] Together, we coded 81.72% ($n = 219$) of the studies in our sample for the variable of temporal precision. Based on the coding scheme, half of the studies were coded as "0" (45.15%; $n = 121$) and the remainder were coded as "1" (20.90%; $n = 56$), "2" (10.82%; $n = 29$), or "3" (4.85%; $n = 13$). The initial agreement rate was 87% and disagreements were discussed and resolved. The initial Cohen's Kappa value for this coding was 0.74.

Second, some studies were explicit about their change trigger, such as attributing a change to newcomer socialization (McNatt & Judge, 2004), or organizational change (Petrou et al., 2018). When a change trigger was either explicitly described or able to be inferred from the authors' description, we coded "1" ($k = 150$). For studies in which no reasons were given to expect a change, we coded "0" ($k = 118$). This is a dummy code and the referent category for the change trigger is "no change specified." The initial agreement rate was 91%, and the initial Cohen's Kappa value was 0.81. Disagreements were again resolved through discussion.

Third, frequency was also coded at the study level by coding the number of measurements in a study and then dividing by the total duration of the study in months. The initial agreement rate was 100%, and the Cohen's Kappa value was 1. Using the job performance example above, given that the number of measurements is three and the duration of the study is 4 months, frequency is calculated as 0.75 instances per month. Overall, our sample included 22 samples at 0.083 instances per month (∼once per year) or fewer; 145 samples with frequencies between 0.083 and 1 instances per month; 54 samples with frequencies between 1 and 4 instances per month (i.e., ∼once per week); and 47 samples with a frequency greater than 4 instances per month.

3.3 | Meta-analytic technique

Because our data are nested, our analytic techniques must account for potential correlated errors, which might exist because (1) multiple measurements were taken for the same variable over time, and (2) multiple variables were measured in the same study. Importantly, meta-analysis is a special case of multilevel modeling, in that sample effect sizes can be viewed as nested within the population of studies (Erez et al., 1996; Gooty et al., 2021; Hox & Leeuw, 2003). We used restricted maximum likelihood estimation for these analyses (Kreft & De Leeuw, 1998). In modeling observed change, we used a random-effects model that assumes that the true effect could vary from study

[2] Our results from dummy coding are consistent with those from effect coding, and the findings are available upon request.

[3] The theoretical rationale of Hypothesis 1 is that a greater level of temporal precision in a study will lead to greater observed change. Please note that this prediction assumes that longitudinal studies are designed to detect change. However, we found that 2.61% of the primary studies were theoretically interested in capturing variability or stability (see Li et al., 2016, for an example). When this type of study sets out to capture longitudinal variability or stability with a greater level of temporal precision, it should capture less observed change. Given that these studies are not consistent with our theoretical rationale and the vast majority of longitudinal studies in our field, we thus excluded them from the test of the temporal precision hypothesis.

to study, and sampling error is not the only reason for the observed differences (Gonzalez-Mulé & Aguinis, 2018). The tests of our hypotheses are based on mixed-effects models, which add fixed effects to explain between-study variance, or heterogeneity, in the population of studies (i.e., temporal precision in H1; change trigger in H2; variable type in H3; rating source in H4; frequency in H5; time interval in H6).

Before conducting the analyses, we transformed the raw data into meaningful measures of change. Using the mean and standard deviation values of all the focal variables at different time points, we calculated Cohen's *d* as the observed change between any two time points. We followed this procedure for all the variables and all possible pairs of time points. Then, because we are only interested in the magnitude of change, as opposed to its directionality, we took the absolute value of Cohen's *d* as the functional measure of change. This is because temporal decisions do not imply directionality but only magnitude. For example, if a time interval is able to detect positive changes over time, it is also able to detect negative changes over time. In performing the subsequent analyses, we transformed Cohen's *d* into a correlation coefficient *r*, which was then transformed into a Fisher's *Z*. The transformation formula between Cohen's *d* and Pearson correlation *r* is

$$r = \frac{d}{\sqrt{d^2 + a}}$$

where *a* is a correction factor for cases where $n_1 \neq n_2$, $a = \frac{(n_1+n_2)^2}{n_1 n_2}$. In the vast majority of cases in our sample of studies, n_1 was equal to n_2, in which case a is simply a constant "4." We then transformed *r* into Fisher's *Z* with the formula below,

$$\text{Fisher's}\, Z = 0.5 * (log(1+r) - log(1-r))$$

Researchers advocating the use of a multilevel approach to meta-analyses suggest using Fisher's *Z* for the effect size (Erez et al., 1996; Hox & Leeuw, 2003) as doing so helps avoid issues that can arise when the effect size correlates with its sampling variance and relaxes the assumption of the effect sizes being normally distributed (Borenstein et al., 2009). We thus conducted all the analyses using Fisher's *Z* and then transformed Fisher's *Z* back to Cohen's *d*, reversing the formulas above to solve for *r* and *d*, respectively, to provide a more intuitive interpretation of our findings.

As we noted earlier, our data are nested at three levels, because any particular longitudinal study (i.e., Level 3) reports information on various variables (i.e., Level 2), which are in turn measured on multiple occasions; the magnitude of the difference in the variables between measurement occasions is our outcome of interest (i.e., Level 1). Thus, we construct three-level, variance-known models, as is required by random-effects meta-analysis (Erez et al., 1996; Raudenbush & Bryk, 2002; Sturman et al., 2005). We computed the Level 1 sampling variance (σ^2) of Fisher's *Z* as $\sigma 2 = 1/(N_k - 3)$, where N_k is the study sample size of the k^{th} sample or study (Hox & Leeuw, 2003; Raudenbush & Bryk, 2002). To test our hypotheses, we began by estimating a random-effects-only model. The null model in Equation 1 provides an estimate of the grand mean of the effect sizes and accounts for sampling variance at Level 1 (i.e., the effect size level), variance at Level 2 (i.e., variable level), and variance at Level 3 (i.e., study level). It is analogous to an estimate of the overall correlation, corrected for sampling error, in traditional meta-analysis. Formally the equation is

$$\begin{aligned} \text{Observed change} \quad ijk &= \pi_{0jk} + e_{ijk}, \\ \pi_{0jk} &= \beta_{00k} + r_{0jk}, \text{ and} \\ \beta_{00k} &= \gamma_{000} + u_{00k}. \end{aligned} \tag{1}$$

The subscripts represent the i^{th} effect size of the j^{th} variable in the k^{th} study. According to these equations, the observed change (Observed change$_{ijk}$) is equal to the true (or average) change (π_{0jk}) and between-study variance $\left(e_{ijk}\right)$. In the Level 2 (i.e., variable level) equation, the π coefficient at Level 1 for thej^{th} variable in the k^{th} study is treated as an outcome predicted by the Level 2 coefficient (i.e., β_{00k}) and the Level 2 error term (i.e., r_{0jk}). In the Level 3 (i.e., study level) equation, theβ coefficient at Level 2 from the k^{th} study is treated as an outcome predicted by the Level 3 coefficient (i.e., γ_{000}) and the Level 3 error term (i.e., u_{00k}).

Following the estimation of our null model, we proceeded by estimating mixed-effects models that incorporate both fixed effects and random effects at different levels. To give one example here, we estimated the following model to test whether the presence of a change trigger affects the observed change:

$$\begin{aligned} \text{Observed change} \quad ijk &= \pi_{0jk} + e_{ijk}, \\ \pi_{0jk} &= \beta_{00k} + r_{0jk} \\ \beta_{00k} &= \gamma_{000} + \gamma_{001}\, \text{Change Trigger} + u_{00k}. \end{aligned} \tag{2}$$

In this model, the observed change is a function of the Level 2 intercept (β_{00k}) and error term (r_{0jk}). The Level 2 intercept is, in turn, a function of the Level 3 intercept (γ_{000}), regression coefficient associated with the change trigger (γ_{001}), and error term (u_{00k}). The significance and magnitude of the coefficient associated with the change trigger indicates how strongly the presence of a change trigger affects the observed change of the i^{th} effect size of the j^{th} variable in the k^{th} study. Other models follow a similar rationale.

4 | RESULTS

Table 1 contains means, standard deviations, and intercorrelations among all variables at three different levels: effect size level (i.e., observed change and time interval), variable level (i.e., variable type and rating source), and study level (i.e., temporal precision, change trigger, and frequency). These statistics are all raw scores before sampling variance weighting. Among the 4812 effect sizes, the time interval has a mean of over 1 year (*M* = 16.04 months) with high dispersion (*SD* = 25.73). Among the 873 variable types, most represent attitudes (46.51%) and are self-reports (78.01%). Among the 268 studies, 55.97% specified a change trigger (*n* = 150), 36.57% had

TABLE 1 Descriptive statistics and intercorrelations.

			1		2		3		4		5		6		7		8		9	
	M	*SD*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Effect size level (*k* = 4812)																				
1. Observed change	0.12	0.17																		
2. Time interval	16.04	25.73	−.09	*.000*																
Variable level (*k* = 873)																				
1. Performance	0.11	0.32																		
2. Non-perf. Behavior	0.23	0.42	−.19	*.000*																
3. Personal attribute	0.07	0.25	−.10	*.004*	−.15	*.000*														
4. Attitude	0.47	0.50	−.33	*.000*	−.51	*.000*	−.26	*.000*												
5. Emotion	0.06	0.24	−.09	*.008*	−.14	*.000*	−.07	*.043*	−.24	*.000*										
6. Self	0.78	0.41	−.52	*.000*	−.14	*.000*	.09	*.008*	.42	*.000*	.12	*.000*								
7. Supervisor	0.03	0.18	.42	*.000*	−.06	*.107*	−.05	*.137*	−.15	*.000*	−.05	*.168*	−.35	*.000*						
8. Subordinate	0.04	0.19	.05	*.178*	.22	*.000*	.02	*.569*	−.17	*.000*	−.05	*.147*	−.37	*.000*	−.04	*.286*				
9. Peer	0.01	0.12	.02	*.644*	.07	*.042*	−.03	*.324*	−.10	*.005*	−.03	*.361*	−.23	*.000*	−.02	*.501*	−.02	*.479*		
10. Objective	0.12	0.32	.39	*.000*	.06	*.084*	−.07	*.037*	−.32	*.000*	−.08	*.024*	−.69	*.000*	−.07	*.046*	−.07	*.036*	−.05	*.187*
Study level (*k* = 268)																				
1. Temporal precision	0.63	0.95																		
2. Change trigger	0.56	0.50	.06	*.361*																
3. Frequency	32.18	95.13	−.18	*.007*	.05	*.461*														

Note: Results in this section are raw correlations before sampling variance weighting.

at least one temporally precise hypothesis ($n = 98$), and on average, the frequency of measurement was 32.18 instances per month.[4]

We also calculated additional descriptive analyses to summarize the existing practices of longitudinal research. First, we found that the average time interval to capture personal attributes [mean (m) $= 24.44$ months; standard deviation (sd) $= 33.60$] was the longest, where the time intervals for performance ($m = 10.10$; $sd = 12.53$), non-performance behaviors ($m = 8.07$; $sd = 20.37$), attitudes ($m = 14.97$; $sd = 24.39$), and emotions ($m = 20.47$; $sd = 27.94$) were substantially shorter. Second, we found that 91.80% of personal attributes, 96.55% of attitudes, and 98.08% of emotions were rated by the self. Roughly half of the performance variables were objective measures (46.46%), whereas 18.18% were self-reported, 24.24% were supervisor-rated, and the rest of the variables were rated by peers (2.02%) or subordinates (6.06%). With regard to non-performance behaviors, the majority of these variables were reported by the self (67.17%). In contrast, 15.15% of them were objective measures, whereas 11.11% were rated by subordinates, 3.03% by peers, and 1.52% by supervisor.

4.1 | Hypothesis testing

We next use meta-analytic techniques to test the hypotheses. The findings of hypothesis tests are reported in Table 2. To facilitate the interpretation of our findings, we converted the estimates from Table 2 (which are in the Fisher's Z metric) into the Cohen's d metric and reported these estimates and their accompanying confidence intervals (CIs) in Table 3. Specifically, we present the amount of observed change based on the commonly used values of the independent variables. That is, because researchers conducting longitudinal studies may wish to know the amount of observed change detected by similarly designed studies in the past, we choose to focus on commonly used time intervals in longitudinal research, including 1 day, 1 week, 1 month, 2 months, 6 months, 1 year, 3 years, and 5 years. In order to test whether $\bar{d}$ scores were significantly different from each other within each category (instead of only in relation to the referent category), we ran a series of k models, where k is the number of subcategories within a category, in which we changed the referent to correspond to each variable. For example, for the variable type category, there are five main subcategories, so we ran five models with a different subcategory serving as the referent in each model. Taken together, these models test whether the differences between each subcategory are significantly different from one another. These results are reported in Table 3, with the results of statistical significance tests denoted using superscripts. As we noted previously, we first ran a null model, the result of which shows the weighted mean observed change across all the effect sizes in our study is Fisher's $\bar{Z} = 0.13$ ($p = 0.000$), equivalent to a $\bar{d}$ of 0.26 (95% CI [0.22, 0.30]). Further, there was significant heterogeneity in the sample of studies ($T^2 = 0.01$; $p = 0.000$) suggesting that there are variable- and study-level factors that contribute to varying levels of change reported across the studies in our database.

Hypothesis 1 argued that greater temporal precision across a study's research hypotheses should lead to greater observed change. We found that temporal precision is positively related to the observed change in a study ($\gamma = 0.01$; $p = 0.047$). As reported in Table 3, a study with hypotheses that met all three criteria (i.e., timing, duration, and shape) demonstrated the highest level of change ($\bar{d} = 0.28$; 95% CI [0.24, 0.32]), followed by studies with decreasing levels of precision (i.e., for studies coded as “2” $\bar{d} = 0.26$; 95% CI [0.22, 0.30], “1” $\bar{d} = 0.24$; 95% CI [0.20, 0.28], and “0”($\bar{d} = 0.22$; 95% CI [0.18, 0.26]). Thus, Hypothesis 1 is supported.

Hypothesis 2 argued that the observed change will be greater in studies with a specified change trigger. As shown in Table 2, we found that the presence of a change trigger is positively related to the observed change ($\gamma = 0.06$; $p = 0.000$). As shown in Table 3, studies that reported a specific change trigger had a $\bar{d}$ of 0.32 (95% CI [0.24, 0.40]), whereas those that did not had a $\bar{d}$ of 0.20 (95% CI [0.12, 0.28]). As a result, Hypothesis 2 is supported.

Hypothesis 3 argued that observed change differs across variable types, following the ascending order of stability from personal attributes (least change), performance, non-performance behaviors, attitudes, to emotions (most change). The effect coding results presented in Table 2 suggest that the changes of personal attributes ($\gamma = -0.04$; $p = 0.001$) and emotion ($\gamma = -0.04$; $p = 0.012$) were significantly below the grand mean, while the change of performance ($\gamma = 0.05$; $p = 0.000$) was significantly above the grand mean. Consistent with our expectation, personal attributes demonstrated the lowest level of change ($\bar{d} = 0.10$; 95% CI [0.02, 0.18]). However, we found that performance showed the highest level of change ($\bar{d} = 0.28$; 95% CI [0.20, 0.36]), followed by non-performance behaviors ($\bar{d} = 0.16$; 95% CI [0.08, 0.24]), attitudes ($\bar{d} = 0.16$; 95% CI [0.12, 0.20]), and emotions ($\bar{d} = 0.10$; 95% CI [0.02, 0.18]). A surprising finding is that emotions and personal attributes had similar levels of stability over time. We thus conducted a post hoc analysis and confirmed that the change in performance and emotions did not significantly differ from each other ($\gamma = 0.00$; $p = 0.847$). We speculate that this is perhaps because emotions, due to their momentary nature, are different from other variable types such that longer time intervals are appropriate for other variable types but not for emotions. We thus conducted another post hoc analysis and found that, with the exception of emotions, all other variable types had greater observed change when measured with longer time intervals (such as performance, $\gamma = 0.004$; $p = 0.000$). However, given that not all variable types followed the order of our prediction, Hypothesis 3 only received partial support.

Hypothesis 4 argued that observed change varies across rating sources, following the ascending order of stability from subordinate rating (least change), peer rating, supervisory rating, self-rating, and objective measure (most change). As shown in Table 2, the means for supervisor- ($\gamma = -0.06$; $p = 0.002$) and self-rated ($\gamma = -0.03$; $p = 0.009$) variables were significantly below the grand mean, while the mean for objectively measured variables ($\gamma = 0.07$; $p = 0.000$)

[4] Please note that the mean score drops to 2.57 after removing 25 outliers, which we address in a post hoc analysis.

TABLE 2 Meta-analytic results.

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	γ (SE)	*p*	γ (SE)	*p*	γ (SE)	*p*	γ (SE)	*p*	γ (SE)	*p*	γ (SE)	*p*	γ (SE)	*p*
Intercept	.13 (.01)	*.000*	.11 (.01)	*.000*	.10 (.01)	*.000*	.13 (.01)	*.000*	.15 (.01)	*.000*	.13 (.01)	*.000*	.12 (.01)	*.000*
Hypothesis 1														
Temporal precision			.01 (.01)	*.047*										
Hypothesis 2														
Change trigger					.06 (.02)	*.000*								
Hypothesis 3														
Personal attribute							−.04 (.01)	*.001*						
Performance							.05 (.01)	*.000*						
Non-perf. Behavior							−.01 (.01)	*.489*						
Attitude							−.01 (.01)	*.093*						
Emotion							−.04 (.01)	*.012*						
Hypothesis 4														
Subordinate									.01 (.02)	*.538*				
Peer									−.01 (.03)	*.806*				
Supervisor									−.06 (.02)	*.002*				
Self									−.03 (.01)	*.009*				
Objective measure									.07 (.01)	*.000*				
Hypothesis 5														
Frequency											.00 (.00)	*.000*		
Hypothesis 6														
Time interval													.002 (.00)	*.000*
The Square Term of Time Interval													−.00 (.00)	*.000*

Note: Coefficients are presented in the table with robust standard errors in parentheses. Coefficients are based on observed amount of change in Fisher's *Z* metric. Sample sizes are $k = 4812$ at the effect size level, $k = 873$ at the variable level, and $k = 268$ at the study level for all models.

TABLE 3 Estimates of the observed change at different levels of the study variables.

Variable	*k*	$\bar{d}$	95% CI
Overall change	4812	0.26	0.22, 0.30
Temporal precision			
a. Hypotheses met zero of the three criteria (timing, duration, or shape)	121	0.22	0.18, 0.26
b. Hypotheses met one of the three criteria	56	0.24	0.20, 0.28
c. Hypotheses met two of the three criteria	29	0.26	0.22, 0.30
d. Hypotheses met all three criteria	13	0.28	0.24, 0.32
Change trigger			
a. No trigger ^{b}	118	0.20	0.12, 0.28
b. Trigger specified ^{a}	150	0.32	0.24, 0.40
Variable type			
a. Personal attribute $^{b\ c\ d}$	61	0.10	0.02, 0.18
b. Performance $^{a\ e}$	99	0.28	0.20, 0.36
c. Non-perf. Behavior $^{a\ e}$	198	0.16	0.08, 0.24
d. Attitude $^{a\ e}$	406	0.16	0.12, 0.20
e. Emotion $^{b\ c\ d}$	52	0.10	0.02, 0.18
Rating source			
a. Subordinate $^{b\ c\ d}$	32	0.34	0.26, 0.42
b. Peer $^{a\ e}$	13	0.30	0.14, 0.46
c. Supervisor $^{a\ e}$	29	0.20	0.08, 0.32
d. Self $^{a\ e}$	681	0.26	0.18, 0.34
e. Objective $^{b\ c\ d}$	102	0.46	0.38, 0.54
Frequency			
a. Once every 5 years $^{f\ g}$	3	0.26	0.26, 0.26
b. Once every 3 years $^{f\ g}$	5	0.26	0.26, 0.26
c. Once a year $^{f\ g}$	14	0.26	0.26, 0.26
d. Once every 6 months $^{f\ g}$	66	0.26	0.26, 0.26
e. Once every 2 months $^{f\ g}$	50	0.26	0.26, 0.26
f. Once a month $^{f\ g}$	29	0.26	0.26, 0.26
g. Once every week $^{a\ b\ c\ d\ e\ g}$	54	0.27	0.27, 0.27
h. Once a day $^{a\ b\ c\ d\ e\ f}$	47	0.42	0.42, 0.42
Time interval			
a. 1 day $^{f\ g}$	350	0.19	0.19, 0.19
b. 1 week $^{f\ g}$	19	0.19	0.19, 0.19
c. 1 month ^{g}	476	0.19	0.19, 0.19
d. 2 months ^{g}	475	0.19	0.19, 0.19
e. 6 months ^{g}	1003	0.21	0.21, 0.21
f. 1 year ^{g}	564	0.23	0.23, 0.23
g. 3 years $^{a\ b}$	1164	0.33	0.33, 0.33
h. 5 years $^{a\ b\ c\ d\ e}$	761	0.43	0.43, 0.43

Note: Superscripts in dummy codes denote the statistical significance at the .05 level of comparisons between effect sizes within each categorical variable group and are ordered according to the variable order (e.g., ^{a} refers to the first variable within each group, ^{b} to the second, etc.). The column *k* represents the number of effect sizes for the given value or range.

was significantly above the grand mean. To better understand this finding, we similarly calculated Cohen's *d* and associated CIs for all the rating sources. Consistent with our predictions, we found that objectively rated variables have the largest change ($\bar{d} = 0.46$; 95% CI [0.38, 0.54]). However, the subjective ratings did not exactly follow the proposed order. We found that the highest amount of change was found in subordinate-rated variables ($\bar{d} = 0.34$; 95% CI [0.26, 0.42]), followed by peer- ($\bar{d} = 0.30$; 95% CI [0.14, 0.46]), self- ($\bar{d} = 0.26$; 95% CI [0.18, 0.34]), and supervisor-rated variables ($\bar{d} = 0.20$; 95% CI [0.08, 0.32]). As a post hoc analysis, we combined self-, supervisor-, peer- and subordinate-ratings into one single subjective rating category and then compared it with objective measures. By using this method, we found that the amount of change for objectively measured variables was significantly larger than subjectively measured variables ($\gamma = 0.09$; $p = 0.000$). However, although we found that objectively measured variables produced more change than all other subjectively measured variables, the differences among subjectively measured variables did not exactly follow the order we predicted. Thus, Hypothesis 4 only received partial support.

Hypothesis 5 argued that frequency should be positively related to observed change. We found that our sample included some outliers with a study duration of hours or minutes (e.g., Howe, 2019; Jiang et al., 2019; Kapadia & Melwani, 2021; Sitzmann & Ely, 2010), which repeatedly measured the variables of interest within the short duration and fell more than three standard deviations below the mean. These extreme values were outliers in the frequency distribution (i.e., more than five standard deviations away from the mean). Following the tradition of meta-analyses, we performed an "in and out" analysis and found that the effect of frequency was not significant when including these outliers ($\gamma = 0.000$; $p = 0.455$) but was significant after removing them ($\gamma = 0.002$; $p = 0.000$). These outliers might have clouded how frequency influenced observed change, and thus, we decided to exclude them for this hypothesis test. From the table, we can see that the general pattern is that frequent designs (e.g., daily and weekly frequency) exhibited more observed change than infrequent designs (e.g., monthly and yearly frequency). We also performed supplementary analyses for frequency scores that are three standard deviations above the mean. The findings indicate that the results remained substantively identical when omitting primary studies that were outliers. Thus, Hypothesis 5 is supported.

Lastly, Hypothesis 6 predicted that time interval should be related to observed change in a curvilinear, inverted-U pattern. As expected, we found that the squared term of time interval was small in magnitude but statistically significant ($\gamma = -0.000$; $p = 0.000$). We also computed the observed amount of change at the most commonly used time intervals in the primary studies (i.e., 1 day, 1 week, 1 month, 2 months, 6 months, 1 year, 3 years, and 5 years). As shown in Table 3, the observed change rises from $\bar{d} = 0.19$ (95% CI [0.19, 0.19]) at a time interval of 1 day, up to $\bar{d} = 0.23$ (95% CI [0.23, 0.23]) at a time interval of 1 year, and then a $\bar{d} = 0.43$ (95% CI [0.43, 0.43]) at 5 years. As a post hoc analysis, we calculated the inflection point of the curvature and found it to be 137.93 months (i.e., ~11.49 years), meaning that if time interval was shorter than 137.93 months, the trend of the relationship was upward ($slope = 0.002$; $p = 0.000$) but if longer than 137.93 months, the trend turned downward ($slope = -0.001$; $p = 0.003$). Although we did not predict a specific time frame for the inflection point of time interval, we do note that it is quite lengthy compared to the majority of our studies that measured in smaller time intervals. Given that the relationship between time interval and observed change was shaped like an inverted-U pattern over time, we claim support for Hypothesis 6. Furthermore, we also tested the interaction effect between frequency and time interval and found that it was statistically significant ($\gamma = 0.03$; $p = 0.001$). When frequency is higher, a longer time interval allows researchers to detect a larger amount of change. When frequency is lower, however, a shorter time interval is more appropriate. The implication of these findings is that researchers must plan enough measurement occasions but also plan enough time to allow change between the measurements.

In sum, among the six hypotheses on (1) temporal precision, (2) change trigger, (3) variable type, (4) rating source, (5) frequency, and (6) time interval, we found full support for Hypotheses 1, 2, 5, and 6 and marginal support for Hypotheses 3 and 4.

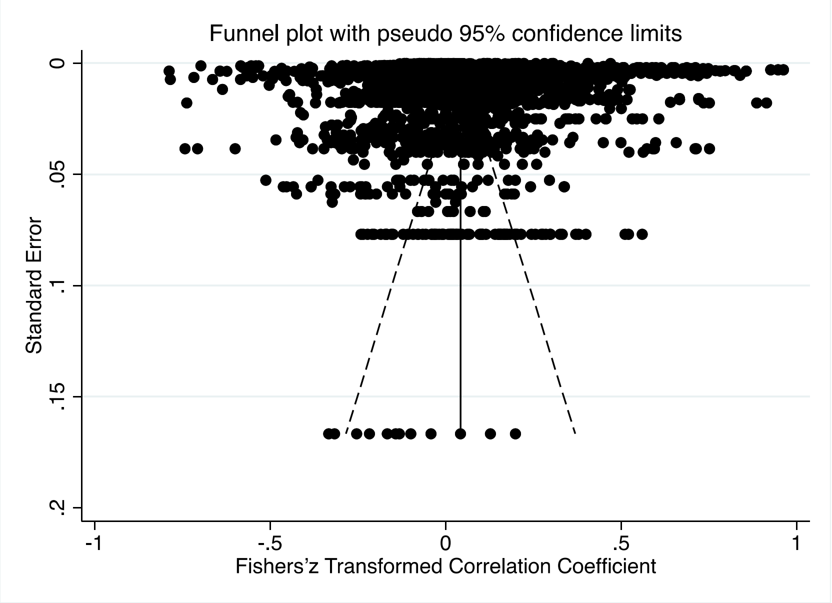
4.2 | Supplementary analyses

We conducted three supplementary analyses to further probe our findings. First, we found that longitudinal researchers have adopted a wide range of analytic techniques. Based on the frameworks developed by Dishop et al. (2020), and Voelkle and Oud (2015), we grouped analytic techniques into two categories. The first category (e.g., latent growth modeling and random coefficient modeling) is where time is explicitly modeled as a variable, while the second category (e.g., latent change score modeling and autoregression) is where time is implied in the order of measurement instances (e.g., an independent variable measured in the first time point relating to a dependent variable measured in the second time point controlling for the dependent variable in the first time point). We found that 20.52% of the samples model time *explicitly* (coded as 1) and 79.48% of the samples model time *implicitly* (coded as 0). We then tested the effect of modeling time explicitly and found that it was positively related to the observed change ($\gamma = 0.05$; $p = 0.018$). As a result, analytic models that model time explicitly exhibit greater observed change than analytic models that model time implicitly.

Second, we conducted a publication bias test using Duval and Tweedie's (2000) trim-and-fill method, implemented in the *metafor* program in R to examine whether the distribution of effect sizes is symmetrical. We conducted these analyses before computing the absolute values of the effect sizes so that both sides of the distribution (i.e., negative and positive change) are represented. The trim-and-fill method allows us to examine the funnel plot of the effect sizes (plotted against their standard errors) to determine whether our database suffers from publication bias (i.e., the suppression of small effects). The funnel plot is shown in Figure 2. It shows that our distribution of effect sizes is almost perfectly symmetrical, as the method did not impute any effect sizes on either side of the distribution. Therefore, we concluded that it was unlikely that publication bias meaningfully affected our study results.

Third, given the longstanding tradition in meta-analyses in the field to correct effect sizes for unreliability, we applied psychometric corrections to the effect sizes in our data and examined the extent to which they differed from the uncorrected effect sizes (Hunter & Schmidt, 1990). Because of the high alphas reported in the primary studies, the uncorrected effect sizes were very similar to the

FIGURE 2 A funnel plot of effect sizes for publication bias test.



corrected effect sizes. Thus, our findings seem to be accurate for typical levels of reliability found in the literature.[5] However, we note that these findings should be interpreted with caution for two reasons. First, the method does not account for effect size dependencies because psychometric corrections are not yet incorporated into multilevel meta-analytic frameworks; there is a need for future research to determine the appropriateness of doing so. Second, the method dichotomizes the continuous moderators, such as temporal precision, time interval, and frequency, resulting in slight differences in the results compared to our multilevel meta-analytic framework.

5 | DISCUSSION

Time plays an important role in theory development and research design because it allows researchers to capture change in a variable of interest. Yet we know little about the relationship between temporal decisions and the change a researcher observes. With a humble intention to explore this uncharted field, we used meta-analytic techniques to examine the effects of a series of temporal decisions. Interestingly, despite a wide search for longitudinal studies in the field, we found that out of the total of 7380 publications in the eight journals we searched, only 204 articles (2.76%) with 268 samples met our criteria for true longitudinal studies (e.g., three or more repeated measures). However, the majority of these articles were published after 2000 (168 articles constituting 82.35% of our sample), confirming the good news that recent studies have been engaging in more longitudinal research as compared to earlier decades (Shipp & Cole, 2015). Yet, as our results demonstrate, the field still needs improvement in how we handle temporal decisions. Our results clearly indicate that the choices researchers make about longitudinal research (i.e., hypotheses, samples, variables, and measurement occasions) affect the observed amount of change they may find. We found that observed change was greater when (1) a study's hypotheses had a higher level of temporal precision; (2) a change trigger was specified; (3) objective measures of variables were used; (4) the variable is a performance variable; (5) the frequency of measurement was greater; and (6) the time interval between measurement instances was longer. Given that observing change is the impetus of studying phenomena over time, these findings offer several important theoretical and methodological implications for longitudinal research in our field.

5.1 | Theoretical and methodological implications

Our meta-analysis contributes to burgeoning longitudinal research by offering informed guidance researchers can use to both deepen the theoretical development of their hypotheses as well as to make informed methodological decisions about temporal design. At a general level, our meta-analytic study offers theoretical contributions by demonstrating the critical nature of providing the theoretical rationale behind temporal decisions. Of particular interest is the fact that, when we screened and coded primary studies, many of the studies did not sufficiently describe the temporal elements of their hypotheses or research settings despite being "true longitudinal research" published in highly ranked academic outlets. Our results indicate that beyond simply designing a strong methodological approach to a longitudinal study, scholars must specify the theoretical rationale for why they expect change and the form it should take within a particular sample.

Further, when attempting to hypothesize and test longitudinal variability, we recommend that future research should report research predictions and findings in a common format. This format would include reporting not only means, standard deviations, and correlation coefficients across all time points but also the precise form of change

[5] These findings are available upon request.

(i.e., temporal precision), why the sample is expected to change (e.g., a trigger or an intervention), the types of variables being studied (e.g., personal attributes or performance), the rating sources of the variables (e.g., objective measures or supervisor ratings), the duration the number of measurement instances (e.g., frequency), and the length of the time interval(s) between measurement instances. We also found that the majority of the longitudinal studies we coded did not report temporal consistency of the repeated measures. Temporal consistency, also known as test–retest reliability, is a measure of reliability that assesses the consistency of scores obtained from the same individuals at different points in time. It examines whether the same individuals would receive similar scores on a measure when it is administered on three or more separate measurement occasions. Most of the longitudinal research only reported Cronbach's alpha, which is a measure of internal consistency reliability used to assess the reliability and consistency of a psychometric scale or test. However, we believe that future research should consider reporting temporal consistency as it is highly relevant to longitudinal research and it also allows us to examine the magnitude of the change of variables. A standardized reporting format for all longitudinal studies would make it easier for readers to interpret results and plan their future work by reviewing previous authors' logic about temporal design, as well as study conclusions compared to other published work.

Beyond these general guidelines, our study also provides more specific research implications, which are summarized in Table 4. First, when a longitudinal study offers a higher level of temporal precision in the theoretical development of its hypotheses, it tends to detect a larger amount of change. This is because better theorization of temporal issues in the hypothesis development stage helps researchers make more informed decisions in the research design stage, allowing them to capture a greater amount of observed change (assuming they are theoretically interested in such change). Yet only 4.85% of the studies in our sample offered fully precise temporal hypotheses, with 45.15% offering zero precision (i.e., no temporal rationale in any of the hypotheses). Our findings suggest that future longitudinal researchers should increase the level of temporal precision in their theory development by specifying the timing, duration, and shape within their hypotheses.

Second, we found that theoretically specifying a change trigger substantially increases the magnitude of observed change. In fact, studies that specified a change trigger produced almost twice as much change as those that did not. This sizeable finding suggests that targeting samples in which change is theoretically expected should be a primary goal for all longitudinal researchers. Yet, we found that 44.03% of longitudinal studies did not specify a change trigger at all. This is consistent with Ployhart and Vandenberg's (2010) observation that longitudinal research may rely too much on convenience samples. Our findings indicate that, despite an increasing trend to conduct longitudinal research, this lack of theoretical and methodological correspondence has continued over the last decade, and it potentially restricts the conclusions researchers can make.

TABLE 4 Summary of findings and implications for future research.

Temporal decisions	Key findings	Implications for future research
Hypotheses	Longitudinal studies will capture greater observed change if the study has at least one temporally precise change-related hypothesis.	Researchers should explicitly theorize the timing, duration, and shape of the research hypotheses. A greater understanding of the theoretical phenomena allows researchers to make more informed design decisions.
Samples	Studies with a change trigger exhibit greater observed change (∼2×) than those without.	Researchers should avoid convenience samples and instead choose a sample that is theoretically expected to change (e.g., a sample with an external event or an intervention).
Variables	Personal attributes exhibit the lowest magnitude of change.	As appropriate to their research stream, researchers should not prioritize personal attributes (unless they wish to predict stability).
	Objective measures exhibit more change than subjective ratings.	As appropriate to their research stream, researchers should prioritize objective measures.
Measurement occasions	Observed change does not differ substantially when the measurement frequency is monthly or yearly. However, the positive effect of frequency becomes salient when the measurement frequency is daily or weekly.	As appropriate to their research stream, using daily or weekly designs allows a researcher to capture a greater amount of change.
	The relationship between time interval and change follows an inverted-U curvilinear pattern. The relationship increases until its peak at 11.48 years.	As appropriate to their research stream, researchers should use longer time intervals. One exception is in the case of measuring emotions where shorter time intervals capture greater change.

Third, we found support for our prediction that among all variable types, personal attributes exhibited the lowest level of observed change. However, somewhat surprisingly, emotions seem more stable than we predicted. Similar to our findings, Houben et al.'s (2015) meta-analysis of emotion dynamics also reported a near-zero between-effect size variance ($\sigma^2 = 0.001$), suggesting a very small amount of observed change in emotion over time.[6] This possibly can be explained that emotions often fluctuate around a stable-baseline (i.e., homeostasis principle), which may result in relatively small changes around the mean level over time. Another speculation is that the time intervals used to capture emotions were too long. According to the findings of the post hoc analysis reported earlier, emotions interacted with time intervals such that that observed change was larger if emotions were measured with shorter time intervals (e.g., daily measures). Researchers seeking to capture change in emotion should consider using shorter time intervals.

Fourth, regarding the rating source hypothesis, we found that subjective ratings exhibited less change compared to objective measures. This finding is consistent with Sturman et al. (2005), although it is important to note that their conclusion was based on job performance only, whereas we extend this finding to all variable types. Scholars may wish to reconsider the specific rating source when they are interested in tracking changes over time, giving preference to objective measures where possible. We also speculate that the smaller magnitude of differences among subjective ratings might have been caused by various rating biases (Landy & Farr, 1980; Murphy & Balzer, 1986). Future research might find ways to reduce these rating biases; we expect the observed amount of change may be higher with such efforts.

Fifth, we found that designing a study with frequent measurement allows researchers to observe more change. However, it is important to point out that only at a very high level of frequency does the benefit of frequent measurement become apparent. Additionally, the significant interaction between time interval and frequency suggests that longitudinal research designs should incorporate longer time intervals and higher frequencies to accurately capture changes over time. While this may pose a challenge for scholars who aim to publish quickly, our findings suggest that without adequate time intervals or measurements, the true effects of a phenomenon may not be fully captured.

Finally, we found as expected that time interval had curvilinear effects on observed change. According to our post hoc analysis, greater change was found as the time interval continued to increase until the point of curvature of 11.49 years. Given that most studies in our field focus on time intervals of weeks, months, or (at most) 1 or 2 years, these findings imply that we may not design our studies with a long enough runway to detect substantial change. Although we recommend that the time interval chosen fit the variable type (i.e., studies of emotions should have shorter time intervals whereas studies of other variable types should have longer time intervals), it may be that management research as a whole would benefit from lengthening the time in which we allow change to unfold in any given study.

5.2 | Study limitations and directions for future research

The conclusions of our meta-analysis are subject to a number of limitations. First, meta-analysis has its own limitations because it is not able to account for all of the granular decisions researchers might make in designing longitudinal studies (e.g., the theoretical origin of the primary study, the subtle differences among different types of constructs, and the unique characteristics of the research site). As a result, it is impossible to develop a guide of optimal time design for future research that captures the true form of change for every situation, individual, and organization (Kozlowski, 2009; Zapf et al., 1996). This one-size-fits-all formula may not even exist because the true form of change is unknown to researchers both before and after the data collection and analysis. That said, part of the value of meta-analysis is an ability to summarize the literature at a general level and inform future research investigations across studies (Borenstein et al., 2009; Cooper et al., 2009), which was the aim of this article. Whereas more specific research questions are more appropriate for a primary study, we see value in both lines of research: using meta-analysis to inform broader research questions and provide general guidance to longitudinal researchers, then using primary studies to investigate the ramifications of more granular decisions.

The second limitation is that we made conclusions across studies with different theoretical traditions and widely different measurement customs. Due to the diversity of research streams represented within each category, we are not able to test different effects within specific research streams in our meta-analysis because of the small *k* across specific literatures. Therefore, it is challenging to make conclusions that specifically apply to each different research program. We believe it may be promising to conduct future meta-analyses that are program-specific, looking into longitudinal studies within a particular research stream. For example, we speculate that research on newcomers may have its own unique change relationships, which could differ from research on organizational change.

Third, similar to all meta-analyses, our conclusions are highly dependent on the set of primary studies included. Although we have made best efforts to code a representative and unbiased sample of primary studies, as longitudinal research continues to grow in popularity with studies of longer duration and/or higher frequency, it is plausible that our conclusions may change over time. We recommend future meta-analyses update our findings every decade or so, perhaps with a different sampling strategy (e.g., more journals), to determine how these conclusions may change.

Fourth, we acknowledge that how to theorize time and design studies that appropriately capture change are two of the biggest and most important challenges in longitudinal research. Although our meta-analysis made important progress by addressing how temporal

[6] The primary studies included in this meta-analysis only represented a subsample of emotion studies that measured both psychological well-being and emotions.

choices produce different results, our study was unable to address all unanswered questions. Therefore, in addition to meta-analysis, future research may consider other methods such as qualitative review, algebraic proof, and Monte Carlo Simulation (Cole & Maxwell, 2009; Dormann & Griffin, 2015; Ployhart & Vandenberg, 2010) to further explore ways to resolve this challenge. It also might be promising to use more inductive or abductive approaches to study change phenomena in and across organizations.

Finally, we have made best efforts to code a representative and unbiased sample of primary studies. However, some of the studies were valuable but failed to meet our inclusion criteria and thus were excluded. For example, ESM studies have a high number of measurements and they allow researchers to detect more complex patterns. We encourage future research to meta-analyze the rates of change in ESM studies specifically.

6 | CONCLUSION

Theorizing and designing a longitudinal study to detect change is challenging, and it is helpful to understand how the prior longitudinal research informs the future longitudinal research. In this meta-analysis, we examined how temporal issues related to the hypotheses, samples, variables, and measurement occasions influence the observed change. Our findings indicate that these decisions affect the magnitude of change reported by primary studies, underscoring the importance of making informed temporal decisions in future longitudinal research.

ACKNOWLEDGMENTS

We acknowledge the help of Yue Yuan for her dedicated support as a research assistant for this project.

CONFLICT OF INTEREST STATEMENT

There are no conflicts of interest involved in this paper.

DATA AVAILABILITY STATEMENT

Data available in article supplementary material.

ORCID

Erica Xu https://orcid.org/0000-0002-8074-9358

REFERENCES

(For a full list of studies included in the meta-analysis, see Data S1)

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. (2011). Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management*, *37*(1), 5–38. https://doi.org/10.1177/0149206310377113

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, *84*(5), 888–918. https://doi.org/10.1037/0033-2909.84.5.888

Alkharusi, H. (2012). Categorical variables in regression analysis: A comparison of dummy and effect coding. *International Journal of Education*, 4(2), 202–210. https://doi.org/10.5296/ije.v4i2.1962

Ambrose, M. L., & Cropanzano, R. (2003). A longitudinal analysis of organizational fairness: An examination of reactions to tenure and promotion decisions. *Journal of Applied Psychology*, *88*(2), 266–275. https://doi.org/10.1037/0021-9010.88.2.266

Anderson, N. H., & Butzin, C. A. (1974). Performance = Motivation × Ability: An integration-theoretical analysis. *Journal of Personality and Social Psychology*, *30*(5), 598–604. https://doi.org/10.1037/h0037447

Barclay, L. J., & Kiefer, T. (2019). In the aftermath of unfair events: Understanding the differential effects of anxiety and anger. *Journal of Management*, *45*(5), 1802–1829. https://doi.org/10.1177/0149206317739107

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & MacKenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology*, *48*(3), 587–605. https://doi.org/10.1111/j.1744-6570.1995.tb01772.x

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. https://doi.org/10.1002/9780470743386

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449. https://doi.org/10.1037/a0038047

Cole, D. A., & Maxwell, S. E. (2009). Statistical methods for risk-outcome research: Being sensitive to longitudinal structure. *Annual Review of Clinical Psychology*, *5*, 71–96. https://doi.org/10.1146/annurev-clinpsy-060508-130357

Cooper, H., Hedges, L., & Valentine, J. (2009). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

Cortina, J. M., Aguinis, H., & DeShon, R. P. (2017). Twilight of dawn or of evening? A century of research methods in the Journal of Applied Psychology. *Journal of Applied Psychology*, *102*(3), 274–290. https://doi.org/10.1037/apl0000163

Costa, P., & Mccrae, R. R. (1999). *A five-factor theory of personality*. Guilford Press.

Day, D. V., Sin, H. P., & Chen, T. T. (2004). Assessing the burdens of leadership: Effects of formal leadership roles on individual performance over time. *Personnel Psychology*, *57*(3), 573–605. https://doi.org/10.1111/j.1744-6570.2004.00001.x

Dimotakis, N., Scott, B. A., & Koopman, J. (2011). An experience sampling investigation of workplace interactions, affective states, and employee well-being. *Journal of Organizational Behavior*, *32*(4), 572–588.

Dishop, C. R., Olenick, J., & DeShon, R. P. (2020). Principles for taking a dynamic perspective. In Y. Griep & S. D. Hansen (Eds.), *Handbook on the temporal dynamics of organizational behavior* (pp. 26–43). Edward Elgar Publishing. https://doi.org/10.4337/9781788974387.00010

Dormann, C., & Griffin, M. A. (2015). Optimal time lags in panel studies. *Psychological Methods*, *20*(4), 489–505. https://doi.org/10.1037/met0000041

Douglas, C., & Gardner, W. L. (2004). Transition to self-directed work teams: Implications of transition time and self-monitoring for managers' use of influence tactics. *Journal of Organizational Behavior*, *25*(1), 47–65. https://doi.org/10.1002/job.244

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich.

Eatough, E. M., Meier, L. L., Igic, I., Elfering, A., Spector, P. E., & Semmer, N. K. (2016). You want me to do what? Two daily diary studies of illegitimate tasks and employee well-being. *Journal of Organizational Behavior*, *37*(1), 108–127. https://doi.org/10.1002/job.2032

Ehrlich, H. J. (1969). Attitudes, behavior, and the intervening variables. *The American Sociologist*, *4*(1), 29–34.

Elfenbein, H. A. (2007). Emotion in organizations: A review and theoretical integration. *Academy of Management Annals*, *1*(1), 315–386. https://doi.org/10.5465/07855981210.5465/078559812

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, *37*(7), 1097–1126. https://doi.org/10.1037/0022-3514.37.7.1097

Erez, A., Bloom, M. C., & Wells, M. T. (1996). Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. *Personnel Psychology*, *49*(2), 275–306. https://doi.org/10.1111/j.1744-6570.1996.tb01801.x

Farrell, J. N., & McDaniel, M. A. (2001). The stability of validity coefficients over time: Ackerman's (1988) model and the General Aptitude Test Battery. *Journal of Applied Psychology*, *86*(1), 60–79. https://doi.org/10.1037/0021-9010.86.1.60

Frank, E. L., Matta, F. K., Sabey, T. B., & Rodell, J. B. (2022). What does it cost you to get there? The effects of emotional journeys on daily outcomes. *Journal of Applied Psychology*, *107*(7), 1203–1226. https://doi.org/10.1037/apl0000908

Fritz, C., Sonnentag, S., Spector, P. E., & McInroe, J. A. (2010). The weekend matters: Relationships between stress recovery and affective experiences. *Journal of Organizational Behavior*, *31*(8), 1137–1162. https://doi.org/10.1002/job.672

Fugate, M., Kinicki, A. J., & Scheck, C. L. (2002). Coping with an organizational merger over four stages. *Personnel Psychology*, *55*(4), 905–928. https://doi.org/10.1111/j.1744-6570.2002.tb00134.x

Gonzalez-Mulé, E., & Aguinis, H. (2018). Advancing theory by assessing boundary conditions with metaregression: A critical review and best-practice recommendations. *Journal of Management*, *44*(6), 2246–2273. https://doi.org/10.1177/0149206317710723

Gonzalez-Mulé, E., & Yuan, Z. (2022). Social support at work carries weight: Relations between social support, employees' diurnal cortisol patterns, and body mass index. *Journal of Applied Psychology*, *107*, 2101–2113. https://doi.org/10.1037/apl0000990

Gooty, J., Banks, G. C., Loignon, A. C., Tonidandel, S., & Williams, C. E. (2021). Meta-analyses as a multi-level model. *Organizational Research Methods*, *24*(2), 389–411. https://doi.org/10.1177/109442811985747

Hannan, M. T., & Freeman, J. (1984). Structural inertia and organizational change. *American Sociological Review*, *49*(2), 149–164. https://doi.org/10.2307/2095567

Harris, M. M., Gilbreath, B., & Sunday, J. A. (1998). A longitudinal examination of a merit pay system: Relationships among performance ratings, merit increases, and total pay increases. *Journal of Applied Psychology*, *83*(5), 825–831. https://doi.org/10.1037/0021-9010.83.5.825

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, *41*(1), 43–62. https://doi.org/10.1111/j.1744-6570.1988.tb00631.x

Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology*, *63*(1), 119–151. https://doi.org/10.1111/j.1744-6570.2009.01164.x

Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. *Journal of Applied Psychology*, *63*(5), 579–588. https://doi.org/10.1037/0021-9010.63.5.579

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, *141*(4), 901–930. https://doi.org/10.1037/a0038822

Howe, M. (2019). General mental ability and goal type as antecedents of recurrent adaptive task performance. *Journal of Applied Psychology*, *104*(6), 796–813. https://doi.org/10.1037/apl0000379

Hox, J. J., & Leeuw, E. (2003). Multilevel models for meta-analysis. In S. P. Reise & N. Duan (Eds.), *Multilevel modeling methodological advances issues & applications* (pp. 90–111). Lawrence Erlbaum Associates.

Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology*, *75*(3), 334–349. https://doi.org/10.1037/0021-9010.75.3.334

Jiang, L., Yin, D., & Liu, D. (2019). Can joy buy you money? The impact of the strength, duration, and phases of an entrepreneur's peak displayed joy on funding performance. *Academy of Management Journal*, *62*(6), 1848–1871. https://doi.org/10.5465/amj.2017.1423

Judge, T. A., Cable, D. M., Colbert, A. E., & Rynes, S. L. (2007). What causes a management article to be cited—Article, author, or journal? *Academy of Management Journal*, *50*(3), 491–506. https://doi.org/10.5465/amj.2007.25525577

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, *52*(3), 621–652. https://doi.org/10.1111/j.1744-6570.1999.tb00174

Kammeyer-Mueller, J. D., Wanberg, C. R., Glomb, T. M., & Ahlburg, D. (2005). The role of temporal shifts in turnover processes: It's about time. *Journal of Applied Psychology*, *90*(4), 644–658. https://doi.org/10.1037/0021-9010.90.4.644

Kapadia, C., & Melwani, S. (2021). More tasks, more ideas: The positive spillover effects of multitasking on subsequent creativity. *Journal of Applied Psychology*, *106*(4), 542–559. https://doi.org/10.1037/apl0000506

Kiburz, K. M., Allen, T. D., & French, K. A. (2017). Work-family conflict and mindfulness: Investigating the effectiveness of a brief training intervention. *Journal of Organizational Behavior*, *38*(7), 1016–1037. https://doi.org/10.1002/job.2181

Kozlowski, S. W. (2009). The mission and scope of the Journal of Applied Psychology. *Journal of Applied Psychology*, *94*(1), 1–4. https://doi.org/10.1037/a0014990

Kreft, I. G., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Sage. https://doi.org/10.4135/9781849209366

Kuypers, T., Guenter, H., & van Emmerik, H. (2018). Team turnover and task conflict: A longitudinal study on the moderating effects of collective experience. *Journal of Management*, *44*(4), 1287–1311. https://doi.org/10.1177/0149206315607966

Lance, C. E., Teachout, M. S., & Donnelly, T. M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology*, *77*(4), 437–452. https://doi.org/10.1037/0021-9010.77.4.437

Lance, C. E., Vandenberg, R. J., & Self, R. M. (2000). Latent growth models of individual change: The case of newcomer adjustment. *Organizational Behavior and Human Decision Processes*, *83*(1), 107–140. https://doi.org/10.1006/obhd.2000.2904

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*(1), 72–107. https://doi.org/10.1037/0033-2909.87.1.72

Li, W. D., Stanek, K. C., Zhang, Z., Ones, D. S., & McGue, M. (2016). Are genetic and environmental influences on job satisfaction stable over time? A three-wave longitudinal twin study. *Journal of Applied Psychology*, *101*(11), 1598–1619.

Li, W.-D., Fay, D., Frese, M., Harms, P. D., & Gao, X. Y. (2014). Reciprocal relationship between proactive personality and work characteristics: A latent change score approach. *Journal of Applied Psychology*, *99*(5), 948–963. https://doi.org/10.1037/a0036169

Liao, H., & Chuang, A. (2004). A multilevel investigation of factors influencing employee service performance and customer outcomes. *Academy of Management Journal*, *47*(1), 41–58. https://doi.org/10.2307/20159559

Lim, S., & Tai, K. (2014). Family incivility and job performance: A moderated mediation model of psychological distress and core self-evaluation. *Journal of Applied Psychology*, *99*(2), 351–359. https://doi.org/10.1037/a0034486

Lorinkova, N. M., & Bartol, K. M. (2021). Shared leadership development and team performance: A new look at the dynamics of shared leadership. *Personnel Psychology*, *74*(1), 77–107. https://doi.org/10.1111/peps.12409

McCormick, B. W., Reeves, C. J., Downes, P. E., Li, N., & Ilies, R. (2020). Scientific contributions of within-person research in management: Making the juice worth the squeeze. *Journal of Management*, *46*(2), 321–350. https://doi.org/10.1177/0149206318788435

McGrath, J. E. (1988). *The social psychology of time: New perspectives*. Sage.

McNatt, D. B., & Judge, T. A. (2004). Boundary conditions of the Galatea effect: A field experiment and constructive replication. *Academy of Management Journal*, *47*(4), 550–565. https://doi.org/10.2307/20159601

Meier, L. L., Cho, E., & Dumani, S. (2016). The effect of positive work reflection during leisure time on affective well-being: Results from three diary studies. *Journal of Organizational Behavior*, *37*(2), 255–278. https://doi.org/10.1002/job.2039

Mitchell, T. R., & James, L. R. (2001). Building better theory: Time and the specification of when things happen. *Academy of Management Review*, *26*(4), 530–547. https://doi.org/10.5465/AMR.2001.5393889

Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, rater and level effects in 360-degree performance ratings. *Personnel Psychology*, *51*(3), 557–576. https://doi.org/10.1111/j.1744-6570.1998.tb00251.x

Murphy, K. R., & Balzer, W. K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. *Journal of Applied Psychology*, *71*(1), 39–44. https://doi.org/10.1037/0021-9010.71.1.39

O'Boyle, E., Banks, G. C., Carter, K., Walter, S., & Yuan, Z. (2019). A 20-year review of outcome reporting bias in moderated multiple regression. *Journal of Business and Psychology*, *34*(1), 19–37. https://doi.org/10.1007/s10869-018-9539-8

Oreg, S. (2003). Resistance to change: Developing an individual differences measure. *Journal of Applied Psychology*, *88*(4), 680–693. https://doi.org/10.1037/0021-9010.88.4.680

Petrou, P., Demerouti, E., & Schaufeli, W. B. (2018). Crafting the change: The role of employee job crafting behaviors for successful organizational change. *Journal of Management*, *44*(5), 1766–1792. https://doi.org/10.1177/014920631562496110.1177/0149206315624961

Pitariu, A. H., & Ployhart, R. E. (2010). Explaining change: Theorizing and testing dynamic mediated longitudinal relationships. *Journal of Management*, *36*(2), 405–429. https://doi.org/10.1177/0149206308331096

Ployhart, R. E., & Hakel, M. D. (1998). The substantive nature of performance variability: Predicting interindividual differences in intraindividual performance. *Personnel Psychology*, *51*(4), 859–901. https://doi.org/10.1111/j.1744-6570.1998.tb00744.x

Ployhart, R. E., & Vandenberg, R. J. (2010). Longitudinal research: The theory, design, and analysis of change. *Journal of Management*, *36*(1), 94–120. https://doi.org/10.1177/0149206309352110

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*(1), 1–25. https://doi.org/10.1037/0033-2909.132.1.1

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*(1), 59–82. https://doi.org/10.1146/annurev.psych.52.1.59

Rousseau, D. M., & Fried, Y. (2001). Location, location, location: Contextualizing organizational research. *Journal of Organizational Behavior*, *22*(1), 1–13. https://doi.org/10.1002/job.78

Shipp, A. J., & Cole, M. S. (2015). Time in individual-level organizational studies: What is it, how is it used, and why isn't it exploited more often? *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 237–260. https://doi.org/10.1146/annurev-orgpsych-032414-111245

Shipp, A. J., & Fried, Y. (2014). *Time and work, Volume 1: How time impacts individuals*. Psychology Press.

Sitzmann, T., & Ely, K. (2010). Sometimes you need a reminder: The effects of prompting self-regulation on regulatory processes, learning, and attrition. *Journal of Applied Psychology*, *95*(1), 132–144. https://doi.org/10.1037/a0018080

Strauss, K., & Parker, S. K. (2018). Intervening to enhance proactivity in organizations: Improving the present or changing the future. *Journal of Management*, *44*(3), 1250–1278. https://doi.org/10.1177/0149206315602531

Sturman, M. C., Cheramie, R. A., & Cashen, L. H. (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. *Journal of Applied Psychology*, *90*(2), 269–283. https://doi.org/10.1037/0021-9010.90.2.269

Taris, T. W., & Kompier, M. A. (2014). Cause and effect: Optimizing the designs of longitudinal studies in occupational health psychology. *Work and Stress*, *28*(1), 1–8. https://doi.org/10.1080/02678373.2014.878494

Toker, S., & Biron, M. (2012). Job burnout and depression: Unraveling their temporal relationship and considering the role of physical activity. *Journal of Applied Psychology*, *97*(3), 699–710. https://doi.org/10.1037/a0026914

van der Werff, L., & Buckley, F. (2017). Getting to know you: A longitudinal examination of trust cues and trust development during socialization. *Journal of Management*, *43*(3), 742–770. https://doi.org/10.1177/0149206314543475

Vandenberghe, C., Bentein, K., & Panaccio, A. (2017). Affective commitment to organizations and supervisors and turnover: A role theory perspective. *Journal of Management*, *43*(7), 2090–2117. https://doi.org/10.1177/0149206314559779

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, *90*(1), 108–131. https://doi.org/10.1037/0021-9010.90.1.108

Voelkle, M. C., & Oud, J. H. (2015). Relating latent change score and continuous time models. *Structural Equation Modeling: a Multidisciplinary Journal*, *22*(3), 366–381. https://doi.org/10.1080/10705511.2014.935918

Wang, M. (2007). Profiling retirees in the retirement transition and adjustment process: Examining the longitudinal change patterns of retirees' psychological well-being. *Journal of Applied Psychology*, *92*(2), 455–474. https://doi.org/10.1037/0021-9010.92.2.455

Woehr, D. J., Sheehan, M. K., & Bennett, W. Jr. (2005). Assessing measurement equivalence across rating sources: A multitrait-multirater approach. *Journal of Applied Psychology*, *90*(3), 592–600. https://doi.org/10.1037/0021-9010.90.3.592

Yu, J. J., Downes, P. E., Carter, K. M., & O'Boyle, E. H. (2016). The problem of effect size heterogeneity in meta-analytic structural equation modeling. *Journal of Applied Psychology*, *101*(10), 1457–1473. https://doi.org/10.1037/apl0000141

Zaheer, S., Albert, S., & Zaheer, A. (1999). Time scales and organizational theory. *Academy of Management Review*, *24*(4), 725–741. https://doi.org/10.2307/259351

Zapf, D., Dormann, C., & Frese, M. (1996). Longitudinal studies in organizational stress research: A review of the literature with reference to methodological issues. *Journal of Occupational Health Psychology*, *1*(2), 145–169. https://doi.org/10.1037//1076-8998.1.2.145

AUTHOR BIOGRAPHIES

Helen Hailin Zhao is an associate professor of Management and Strategy from the University of Hong Kong. Her research interests include time, social network, and human resource management.

Abbie J. Shipp is the M. J. Neeley Professor of Management at Texas Christian University. Her research focuses on the psychological and subjective experience of time at work.

Kameron M. Carter is an assistant professor of Management at Old Dominion University's Strome College of Business. Her research examines interpersonal interactions, work design, and research methods.

Erik Gonzalez-Mulé is an Associate Professor of Organizational Behavior and Human Resources at the Kelley School of Business at Indiana University. His research focuses on employee health and stress, team composition, and meta-analytic methods.

Erica Xu is an associate professor from Hong Kong Baptist University. Her research interests include leadership and team dynamics.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zhao, H. H., Shipp, A. J., Carter, K., Gonzalez-Mulé, E., & Xu, E. (2024). Time and change: A meta-analysis of temporal decisions in longitudinal studies. *Journal of Organizational Behavior*, 1–21. https://doi.org/10.1002/job.2771