Occupational Health Science
https://doi.org/10.1007/s41542-022-00122-y

ORIGINAL RESEARCH ARTICLE

An Alpha, Beta and Gamma Approach to Evaluating Occupational Health Organizational Interventions: Learning from the Measurement of Work-Family Conflict Change

Beth A. Livingston[1] · **Shaun Pichler**[2] · **Ellen Ernst Kossek**[3] · **Rebecca J. Thompson**[4] · **Todd Bodner**[5]

Received: 7 September 2021 / Revised: 11 July 2022 / Accepted: 13 July 2022

Abstract
Given the rapid growth of intervention research in the occupational health sciences and related fields (e.g. work-family), we propose that occupational health scientists adopt an "alpha, beta, gamma" change approach when evaluating intervention efficacy. Interventions can affect absolute change in constructs directly (alpha change), changes in the scales used to assess change (beta change) or redefinitions of the construct itself (gamma change). Researchers should consider the extent to which they expect their intervention to affect each type of change and select evaluation approaches accordingly. We illustrate this approach using change data from groups of IT professionals and health care workers participating in the STAR intervention, designed by the Work Family Health Network. STAR was created to effect change in employee work-family conflict via supervisor family-supportive behaviors and schedule control. We hypothesize that it will affect change via all three change approaches—gamma, beta, and alpha. Using assessment techniques from measurement equivalence approaches, we find results consistent with some gamma and beta change in the IT company due to the intervention; our results suggest that not accounting for such change could affect the evaluation of alpha change. We demonstrate that using a tripartite model of change can help researchers more clearly specify intervention change targets and processes. This will enable the assessment of change in a way that has stronger fidelity between the theories used and the outcomes of interest. Our research has implications for how to assess change using a broader change framework, which employs measurement equivalence approaches in order to advance the design and deployment of more effective interventions in occupational settings.

Keywords Work-family · Intervention · Measurement equivalence

Extended author information available on the last page of the article

Published online: 19 August 2022

Although interventions are central to occupational health science, the evaluation of how they affect change across workplace populations and contexts is under-developed theoretically and empirically (e.g., Kossek, 2016; Bodner & Bliese, 2018). Using data from a major U.S. randomized control trial study involving an organizational work-family intervention called STAR (Kossek et al., 2014) as an exemplar, the goal of this paper is to demonstrate how different types of change in response to an intervention can be hypothesized and assessed.

One of the most popular evaluation methods that is used to assess the efficacy of intervention change is a design that compares pre- and post-intervention scores on an outcome of interest (e.g., Kelly et al., 2014; De Boer et al., 2004; Taimela et al., 2008). Multi-item self-report scales are given to participants before and after an intervention, in a control group and a treatment group, and those items are aggregated into a score, which is compared across groups and across time. These scales are often subjected to a validation process, which means that researchers examine the degree to which the items on that scale hang together (e.g., reliability of scale scores, dimensionality of items) and assess what they claim to assess (e.g., construct validity; Bagozzi et al., 1991). In modern psychometric evaluation, researchers use latent variable models, or factor models, in this process—applying exploratory and/or confirmatory factor analysis in their validation and assessment (Cole, 1987).

This approach is often used as gold-standard of intervention evaluation (e.g., Huang et al., 2015). This method of assessing intervention efficacy also provides an opportunity to observe other possible ways that interventions can create change in employee outcomes. In this paper, we use theories of individual and organizational change (Golembiewski et al., 1976) to suggest three types of changes that can be caused by interventions. They are: alpha changes, when interventions result in different scale scores for outcomes; beta changes, when interventions change the way participants understand scale anchors or how an item relates to other items (Millsap & Hartog, 1988); and gamma changes, reflecting a shift in the way participants conceptualize the construct that the scale purports to measure. This framework has been further developed using measurement invariance concepts (e.g., Spurk et al., 2011). In this paper we apply concepts from measurement invariance and latent variable models to assess whether alpha, gamma, or beta change is present in response to an intervention using scale-score measures of constructs in a work-family intervention.

Alpha change as a result of the STAR intervention has been assessed in prior research as an overall shift in the pre and post test scale scores of work-family conflict (e.g., Kelly et al., 2014; Moen et al., 2016). However, alongside this absolute change in work-family conflict (Bodner & Bliese, 2018), we argue that the intervention may also affect gamma and beta change, which is not assessed by comparing changes in scale scores of a construct over time. These alternative mechanisms of change, however, *can* be be assessed through tests of measurement equivalence. Gamma change is a redefinition or reconceptualization of the construct being measured (Golembiewski et al., 1976; Schmitt, 1982). Because the STAR intervention was focused on increasing family-supportive supervision and schedule control (in order to reduce employee self-reports of work-family conflict), it could have also changed how employees conceptualize the *construct* of work-family conflict in general, changing not how much conflict they report on an aggregated measure of the construct (alpha change) but rather what the construct of *“work-family conflict”* means to them.

Alternatively, beta change refers to a recalibration of the scale points used in measures of a construct (Golembiewski et al., 1976; Schmitt, 1982). Employees do not necessarily interpret the *construct* differently, nor is their *level* of “conflict” necessarily different, but their perceptions of the scale response points has changed due to new positive or negative associations that have been built by the intervention (Thompson et al., 1999). In other words, to assess alpha change, we can compare whether *latent variable means of a measure of a construct* are the same across groups, with gamma change, we compare whether the *latent contruct itself* is the same construct across groups, and with beta change, we compare whether *parameters associated with the items of the measure used to assess the construct* are the same across groups.

In this paper, we argue that directly theorizing and assessing gamma and beta change, in addition to alpha change, (Oreg et al., 2012) can have important implications in terms of how intervention results are interpreted (Allen & Martin, 2017). Specifically, we argue that observing differences in scores on measures over time, or on factor means, may be due to beta or gamma change, and not just alpha change, and that other types of change may obscure our ability to observe alpha change. We thus make three specific contributions to the literature on occupational health interventions. First, research on organizational interventions rarely analyze intervention efficacy using methods other than an “alpha change” approach in which they directly compare the observed scores on measures of focal constructs before and after the intervention (which is one way of assessing alpha change). We demonstrate that other types of change captured using measurement equivalence approaches could be important harbingers of intervention effectiveness. By proposing that occupational health scientists adopt an “alpha, beta, gamma” change approach when hypothesizing intervention efficacy and assessing change, our study helps address current shortcomings in the OHP field that mainly focuses on assessing alpha change in outcomes.

Second, we demonstrate how accounting for (or not accounting for) beta and gamma types of change can affect the evaluation of expected alpha change. Although research has demonstrated the importance of establishing measurement equivalence (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000) before comparing groups, we hypothesize that inequivalence can be indicative of change. Thus, it is important to delineate which type of change is expected and why, a priori, and to understand how an intervention may simultaneously affect different types of change. In making this argument, we are suggesting that measurement variance is not a bad thing in and of itself, but, rather, may have been under-utilized as an additional way to hypothesize and evaluate the efficacy of interventions.

Finally, while using principles of measurement equivalence testing as the mechanism to assess beta and gamma change, we specifically integrate scalar equivalence testing—in which we compare the intercepts of items comprising our scale measures—as an indicator of beta change. Most prior work has overlooked differences on item intercepts, which reflects what the score for an item is if the actual level of the underlying latent construct is zero. We demonstrate that where individuals’ responses *start* on a scale (i.e., an item intercept) may also be affected by an intervention, apart from the intervention affecting changes in the construct itself (i.e., lower levels of latent work-family conflict). We then examine this directly using a Multiple Indicators Multiple Causes approach (MIMIC; Muthén, 1989), which is used to examine the simultaneous effect of exogenous causal variables like interventions on the indicators (items) and the latent variable (construct) at the same time. We demonstrate that this is a fruitful area of theorizing around organizational interventions and change.

Finally, we offer important practical contributions by demonstrating how change associated with an intervention can manifest via what is often only considered to be random measurement error, and how a-priori theorizing around specific types of change can impact accuracy and efficacy of the recommendations intervention researchers make to various constituencies. In the sections that follow, we describe alpha, beta, and gamma change in more depth. Next, we link these concepts to methods of assessment using measurement invariance testing. Then, we present specific hypotheses regarding, gamma, beta, and alpha change using the STAR intervention and test them using the methods we propose.

Literature Review and Study Hypotheses

Three types of individual change – alpha, beta and gamma– can be investigated due to exposure to any particular intervention (Golembiewski et al., 1976; Pitts et al., 1996; Schmitt, 1982), and may co-occur to some degree. Alpha change, a change in the level of a construct, which can be assessed via an observed score of a measure of a construct from one time point to another (Millsap & Hartog, 1988; Riordan et al., 2001; Schaubroeck & Green, 1989), is the most commonly assessed change in interventions and randomized controlled trials. However, depending upon the intervention's theory of change (Breuer et al., 2015; De Silva et al., 2014; Reinholz & Andrews, 2020; Zand & Sorensen, 1975), it may well be that testing additional types of change allows for more specific and robust tests of an intervention's efficacy.

Beta change (Schaubroeck & Green, 1989; Schmitt, 1982) is when the measurement scale that respondents use changes in calibration. This occurs, for example, when participants use more or fewer of the scale points for each item (e.g., scale variances shrink or expand across groups) or people change the degree to which they see a particular item as related to the underlying construct (e.g., changes in factor loadings that associate each item to its underlying construct as described by Schaubroeck & Green, 1989).[1] So, after the intervention, a respondent may see a response scale of 1 = strongly disagree to 5 = strongly agree in

[1] As noted by an anonymous reviewer, it is possible that a shift in a factor loading could represent a shift in the definition of the construct itself (gamma change) and not an understanding of a specific item on a scale (beta change). However, following prior research, we believe that shifts in factor loadings are better understood as indicators of changes in measurement calibration than as qualitative differences in the underlying construct itself.

different ways, such that 1 (strongly disagree) and 2 (disagree) now mean the same thing to them. This type of change does not mean that the level of the actual latent construct has changed, but rather that the items and measures used to capture it are being used in different ways over time by participants. As Vandenberg and Self (1993) note, with beta change: "Although a statistically significant difference exists between the two values, this difference possesses no meaning because in reality, the two values are perceptually identical" (pg. 558).

Gamma change is when the definition or understanding of a construct shifts over time in the minds of participants (Schaubroeck & Green, 1989). Comparing the same construct over time to test for changes in the construct (or alpha change) means comparing latent scores on the continuum underlying the construct—looking for a quantitative shift in the same construct. Gamma change renders comparisons of scores on the latent continuum non-comparable because they are *qualitatively* different constructs. For instance, if at Time 1, you interpret the construct of *work-family conflict* as "not having enough time to fulfill both your work and your family responsbilities," but after the intervention, you define it as "your work and your family expect you to to meet different role expectations," this constitutes a qualitative shift in the construct's definition. While both are legitimate interpretations of a "work family conflict" construct, you cannot compare scale scores across time, as scores are referring to different latent constructs altogether.

Interventions could thus affect outcome measures in all of these ways at the same time, whether via actual changes in levels of a construct (alpha change), reconceptualization of constructs (gamma change), or reinterpretations of the measures and items themselves (beta change) (Menard, 1991; Riordan et al., 2001). And while detecting and interpreting alpha change could be confounded by the presence of gamma and beta change, gamma and beta change may also be a purposeful, designed outcome of a change endeavor. Gamma and beta change, however, are rarely included in studies on organizational change or interventions (Riordan et al., 2001).

According to Riordan et al. (2001), only 6% of articles examining change hypothesized gamma or beta change (as opposed to alpha change). But even among this small percentage, none classified it as such. For instance, examinations of the efficacy of training for rater bias in evaluation is suggestive of beta change (Riordan et al., 2001). Organizational culture change interventions may be suggestive of gamma change in that employees are encouraged to reconceptualize and reinterpret attributes of the culture (Riordan et al., 2001). Neither of these assessments are usually framed in terms of Golembiewski's model. Even those studies that raise the possibility that the intervention could have affected beta or gamma change do not test those theoretical arguments directly (see Holman & Axtell, 2016; Logan & Ganster, 2005). For instance, when Logan and Ganster (2005) investigated why there might not have been observed alpha change in their control intervention to reduce stress, interview results suggested that some respondents may have shifted their understanding of the scales used to assess stress in response to learning and understanding stress and control differently. We argue that these sort of explanations related to both gamma and beta change may actually be *expected* in response to an intervention.

Assessing Change using Measurement Invariance Testing

To assess alpha, beta, and gamma change, researchers use analytical approaches from the measurement invariance and confirmatory factor analysis literatures (Chan & Schmitt, 2000; Schmitt, 1982; Vandenberg & Lance, 2000; Vandenberg & Self, 1993). As shown in Table 1, when different types of change are expected, researchers can test for each using principles of measurement invariance testing.

Prior work establishing methods of assessing beta and gamma change has left out the role of scalar/intercept-based invariance (Millsap & Hartog, 1988; Riordan et al., 2001; Vandenberg, 2002; Vandenberg & Self, 1993). Because scalar non-invariance reflects shifts in intercepts that can reflect item bias (Hofmans et al., 2008; Nye et al., 2010), it is important to delineate whether such changes are linked to the intervention. Consistent with Schmitt (1982) and Vandenberg and Self (1993), we conceptualize changes in item intercepts as beta change. Given beta change is a “source of systematic bias affecting the true-score continuum” (Schaubroeck & Green, 1989, pg. 895), not considering changes in intercept terms may mean that beta change is not adequately assessed. Here, we include scalar non-invariance as an indicator of beta change (as shown in Table 1). If item intercepts were affected by an intervention, this would not reflect underlying shifts in the *latent construct* of interest (e.g., a factor mean), but rather in the interpretation of the item or of its scale, regardless of the latent score of the underlying construct. In factor analysis terms, item intercepts reflect the value of a measurement item irrespective of its relationship to the underlying construct or factor.

Golembiewski et al. (1976) suggest that change in response to an intervention can occur via multiple mechanisms which may be captured via measurement invariance testing (Table 1). We argue, as above, that these additional mechanisms of change may be hypothesized a priori based on the underlying theory of change of a specific intervention (Breuer et al., 2015; De Silva et al., 2014; Reinholz & Andrews, 2020; Zand & Sorensen, 1975). Matching the assessment and analyses to the theoretical mechanism is a critical component of evaluating intervention efficacy and generalizing such effects across alternative contexts. Even when alpha (or direct) change is the intervention’s purpose and a strong way to assess its effects, it is important to correct for confounding effects of beta change and to rule-out gamma change in order to accurately interpret alpha change (Chan, 1998). Thus, gamma and beta change should be tested before alpha change (Schmitt, 1982).

We now turn to our examination of change data from a specific work-family intervention, the STAR intervention (Kossek et al., 2014), as a lens to understand these broader occupational health intervention change evaluation issues.

Work-Family Interventions and Individual Change

As Riordan et al. (2001) note, change is a cornerstone of the applied psychology and organizational behavior literatures. Change in this context refers to “employees reinterpreting and revising both the meaning of work as it pertains to a particular

Table 1 The Overlap of Measurement Equivalence and Organizational Change

	Measurement Equivalence Literature	Organizational Change Literature	Analysis
Same factor model across groups	Configural equivalence (Vandenberg & Lance, 2000)	(lack of) Gamma change (Vandenberg & Self, 1993)	Necessary for continued MEI testing; noninvariance may be a sign of gamma change
Equal factor loadings ($\lambda_{111}=\lambda_{211}=\lambda_{311}$)	Metric equivalence (Vandenberg & Lance, 2000)	(lack of) Beta change (Vandenberg & Self, 1993)	Second step in MEI testing; noninvariance may be a sign of beta change
Equal intercepts ($\tau_{111}=\tau_{211}=\tau_{311}$)	Scalar equivalence (Vandenberg & Lance, 2000)	*Not specified*	Third step in MEI testing; noninvariance may be a sign of beta change
Equal factor covariances (($\Psi_{21}=\Psi_{31}=\Psi_{32}$)	Optional: Reliability (Vandenberg & Lance, 2000)	(lack of) Gamma change (Vandenberg & Self, 1993)	Option in MEI testing; noninvariance may be a sign of gamma change
Equal factor variances ($\Psi_{11}=\Psi_{22}=\Psi_{33}$)	*Not specified*	(lack of) Beta change (Vandenberg & Self, 1993)	Option in MEI testing; noninvariance may be a sign of beta change
Equal factor means ($\xi_1=\xi_2$)	Only if configural, metric, and scalar are achieved	(lack of) Alpha change (Schmitt, 1982)	Option in MEI testing; noninvariance is a sign of alpha change
Equal error variances ($\delta_{111}=\delta_{211}=\delta_{311}$)	Optional: Uniqueness equivalence (reliability) (Vandenberg & Lance, 2000)	*Not specified*	Option in MEI testing

Basic Item Score: $X_{ijk}=\tau_{1jk}+\lambda_{1jk}(\xi_{i1})+\delta_{1jk}$

Response X for respondent i on item j at time k organization and a view of themselves as functioning members…", which can be affected by organizational interventions or employees phenomenological processes (Riordan et al., 2001, pg. 52). Unlike many other foci of change interventions focused on individual-level outcomes, work-family focused change requires attention to multiple roles and domains beyond the work role. Measures of inter-role conflict are dependent upon non-work roles that may not be affected as directly by interventions staged in the work domain, as we know that experiences in one domain affect outcomes in that same domain more strongly than outcomes in others (e.g., Judge et al., 2006). We expect that alternative modes of change are more likely to occur when there are alternative sources of variance built into the construct of interest (i.e., work-family conflict is not only affected by *work*, but also by *family*).

We expect that the STAR work-family intervention (Kossek, 2016), which is focused on multiple components—increasing the control that respondents have over their schedules and promoting family friendly behaviors in supervisors—with the purpose of increasing the work-life balance of employees in two types of firms, a long-term care health facility and an information technology firm, will affect change in multiple ways. The intervention was designed to assess alpha change in the latent construct of work-family conflict, such that people in the treatment group will experience less work-family conflict than the control group (Kelly et al., 2014). But we also expect that it may induce gamma change by affecting how people interpret the construct of work-family conflict and that it may induce beta change by affecting how people perceive the items of the measure of work-family conflict that we use.

STAR trained supervisors on the value of demonstrating support for employees' personal lives and to prompt employees to reconsider when and where they work (Hammer et al., 2009, 2011; Kelly et al., 2011, 2014). Work-family conflict (the interrole conflict between the domains of work and family, measured separately using scale measures of the constructs of work-to-family conflict and family-to-work conflict) was a critical primary outcome of interest. Kelly et al. (2014) and Moen et al. (2016) demonstrated that this work-family intervention induced alpha change in work-family conflict (in the IT organization, *Tomo*, but they did not examine its efficacy in the health care company, *Leef*) and in other well-being outcomes (also in Tomo). The intervention was only effective for certain subgroups in the health care organization (Kossek et al., 2019).

Although the practical principles that Kossek et al. (2014) note in their implementation guide focusing on alpha change are comprehensive, we argue that they miss other indicators of change that can be captured by Golembiewski et al's (1976) tripartite model of change management. The STAR intervention can also affect beta and gamma change, and we can evaluate this change using a factor analysis approach, informed by measurement invariance testing (as per Chan & Schmitt, 2000; Schmitt, 1982).

Gamma change A lack of gamma change in a construct establishes that qualitative perceptions of the latent construct itself are stable across time (and across groups). In other words, we ask: will the STAR intervention change whether "work-family conflict" is seen as the same construct after the intervention (gamma change)? Using measurement invariance methods, this can be assessed by establishing whether the overall factor structure of a model including family-to-work and work-to-family conflict fits the data the same pre- and post-intervention, and whether work-to-family and family-to-work conflict covary similarly across time or groups (Golembiewski et al., 1976; Schaubroeck & Green, 1989; Schmitt, 1982). These methods can indicate whether respondents are thinking about different constructs at different time points based on intervention group (Golembiewski et al., 1976; Riordan et al., 2001; Vandenberg & Self, 1993).

In the context of the current study, it could be the case that employees' knowledge about and understanding of the work-family conflict construct before the intervention was limited. The STAR intervention involved supervisor-targeted family-supportive supervisor training and changes in employee perceptions of support for work and family; employee-targeted change around how and where they work, i.e., schedule control; as well as a shift in organizational culture from a focus on "face time" and long work hours to results (Kelly et al., 2014; Moen et al., 2016). Instead of focusing on work-family challenges as an individual issue, the intervention was designed to "change the rules of the game at work" and to affect how work and personal life could be managed more effectively. More specifically, the STAR intervention was designed to "modify the practices, interactions, and social meanings within this workplace" (Kelly et al., 2014, pg. 487).

Interventions designed to change organizations in this way, i.e., more systemically and as related to organizational culture, "seem inherently to address gamma change" (Riordan et al., 2001, pg. 59). In the case of the STAR intervention, although not necessarily planned, gamma change may have occurred. Since family-supportive supervision, schedule control, and a results orientated culture were not explicit in the organizations being studied prior to the intervention, it is feasible that the STAR intervention could have substantively changed employees' conceptualization of "work and family" in general, including work-to-family and family-to-work conflict. Broader changes in one's psychosocial environment can result in gamma change (Golembiewski et al., 1976) including in measures of work-family conflict across time (Pejtersen & Kristensen, 2009). Thus, although the intervention may not have been expressly designed to change people's definition of work-family conflict (gamma change) or to educate employees who did not understand what work-family conflict was, moving from an uninformed perspective to an informed perspective could have resulted in gamma change.

> Hypothesis 1: A work-family intervention will result in gamma changes in work-family conflict, such that the interpretation of the construct will be variant across control and intervention groups.

Beta change Beta change involves a shift in the scaling and measurement for a specific measure of a construct. This can include a reinterpretation of a rating scale, change in the way in which intervals of a scale are interpreted across time, and how the measure itself is used and interpreted (Riordan et al., 2001; Vandenberg & Self, 1993). Beta change can be assessed by assessing by whether there has been a recalibration of intervals of the scale used to measure the conceptual domain (Golembiewski et al., 1976; Schmitt, 1982). For instance, a response of “strongly disagree” on the work-to-family conflict measure included in the STAR intervention at Time 2 could be interpreted the same as a response of “disagree” on the same measure at Time 1, if the intervention made people apply different decision rules to their perceptions of measures and items. In the context of the current study, while there may be a statistically significant mean difference in work-family conflict across time, this is not necessarily exclusively due to a reduction in the *actual* work-family conflict latent construct but also due to a change in the way in which the scale used to assess work-family conflict is interpreted and how item responses are recorded (Riordan et al., 2001).

In data analytic terms, beta change is traditionally interpreted as being present when factor variances change over time (e.g., a 1–5 scale is used in different ways over time, such as with range restriction) or when an observed item in a scale is differently associated with a latent factor representing a construct (factor loadings) over time (as discussed in Schaubroeck & Green, 1989). In this paper, we also propose that changes in item intercepts may reflect beta change (Table 1), as an item intercept is the value of an item when its association with the underlying latent construct is nil. In other words, all of these methods of assessment refer to the *measure* being used, and not the underlying *latent construct* iself. If an intervention affected change in these scale attributes (compared to the control group), this would be indicative of beta change.

The STAR intervention (Kossek et al., 2014) was intended to change how people perceive their responsibilities and the interference between roles, potentially making people within an intervention group more similar on their perceptions of conflict. This is similar to organizational culture interventions, which can narrow the variance in the perception of workplace attributes (factor variance), enhance the connection between certain items and their underlying constructs (factor loadings), or recalibrate the average response to an item by shifting how people generally respond to a scale over time, regardless of the underlying factor score (intercepts).

Changes in scaling of the work-family construct could have occurred in the context of the intervention due to a variety of factors, some of which are related to the STAR intervention itself. The further apart measurements are taken across time, the less likely alpha change alone is attributable to differences in levels of a construct (Marsh & Grayson, 1994). Since work-family conflict was measured six months apart between Time 1 and Time 2 (Moen et al., 2016), time itself could be a reason for beta change in either the control group or the intervention group: Multiple measurements of the same construct can result in habituation and practice effects (Keefer et al., 2013). So, too, can measurement fatigue (Schwartz et al., 2004), which is possible in any large-scale study such as the STAR intervention wherein respondents complete lengthy surveys at each time point in the study. But we argue that, in response to the intervention itself, which is focused on reducing work-family conflict in part by training front-line supervisors to be family-supportive, could have

Table 2 Measures and Items in the Current Study

WFC1	Demands of work interfere with family/personal time
WFC2	Job makes it difficult to fulfill personal responsibilities
WFC3	Things at home do not get done b/c of demands of job
WFC4	Job strain makes it difficult to fulfill fam/personal duties
WFC5	Due to work make changes to fam/personal activities
FWC1	Demands of fam/personal relationships interfere with work
FWC2	Put off things at work because of demands on time at home
FWC3	Things at work not done b/c of demands of fam/personal life
FWC4	Home life interferes with responsibilities at work
FWC5	Family-related strain interferes w/ ability to do job duties

All items asked at all time periods (1 = baseline, 2 = 6-month lag, 3 = 12-month lag, 4 = 18-month lag)

changed the standards by which employees use to assess their own work-family conflict, resulting in beta change (Hammer et al., 2016).

If managers trained to be more family-supportive were helping employees to see their work and their family responsibilities differently, we may see respondents shift their response scales even without experiencing more or less actual conflict. Take item wfc4 in Table 2: “Job strain makes it difficult to fulfill family or personal duties.” An intervention may not change overall levels of the underlying factor/construct of work-family conflict (alpha change), but may decouple perceptions of “job strain” from the concept of conflict (as compared to item wfc1 “Demands of work interfere with family/personal time”). This might make the factor loading for this item less strong in the intervention group than in the control group, an example of beta change. Alternatively, perhaps the intervention may decrease the work-family conflict factor variances compared to the control group, as employees experience a shared intervention experience and shared interpretations of that reality and become more similar over time. Groups and organizations can promote shared understanding and realities via culture or climate (e.g., Bhave et al., 2010). These shared realities for employees in intervention units may result in more compressed factor variances (i.e., more similarity in how they see work-family conflict; beta change) without affecting the level of it (alpha change).

As with research in assessment and scale development on differential item functioning (DIF; Stark et al., 2006), we expect that intervention group can affect overall item scores (intercepts). Scalar equivalence testing requires one to imagine what a person’s response on a scale item would be if the underlying factor were constrained to zero: in other words, what is a default response on this scale, regardless of the actual level of the underlying construct. A classic DIF example is scalar (intercept) non-invariance seen in a three-item measure of depression that includes “I cry a lot”. Crying is related to the construct of depression (factor loading), but women might report higher levels of crying unrelated to the underlying depression construct. They start at a higher intercept than men do. Likewise, with work-family conflict and differential item functioning based on intervention group status, we might see that being in an intervention group changes how employees respond to certain items in a work-family conflict measure that is not related to the actual conflict they experience (alpha change).

Consider the item wfc3, “Things at home do not get done b/c of demands of job.” An intervention to reconceptualize work demands could mean that a respondent would change from a “neutral” to a “disagree” on this item, even if their perceived conflict overall was actually unchanged. Or, perhaps the intervention could have re-focused employees on the importance of their home life, and thus they select “strongly disagree” instead of a “disagree” on this item, even though their overall conflict—the degree to which work has interfered with home/family—is unchanged. This would be observed with intercept non-invariance and would be an example of beta change, as the intervention changed where someone started on a scale (responding with a 3 vs. 2, etc.) without changing the underlying latent work-to-family conflict construct.

Thus, we expect that an intervention will affect beta changes in work-family conflict.

> Hypothesis 2: A work-family intervention will result in beta changes in work-family conflict, such that the scales used to respond to the construct (beta) will shift as a result of the intervention.

Implications of gamma and beta change Prior research on measurement equivalence suggests that non-equivalence can have grave implications for the validity of factor mean comparisons (e.g., Byrne et al., 1989; Chen, 2008; Vandenberg & Lance, 2000). Likewise, the presence of gamma or beta change may affect the interpretation of alpha change, obscuring our ability to perceive the effect of the STAR intervention on work-family conflict changes over time. We expect that, just as not establishing measurement invariance across cultural groups makes it difficult to be sure we are comparing “apples to apples” when determining change or differences across groups (Ryan et al., 1999), the failure to identify beta or gamma change can affect how the overall alpha change as a result of the intervention is interpreted. To accurately interpret alpha change when beta change is detected would require calibration of the work-family conflict measure to ensure consistency across time (Golembiewski et al., 1976; Schmitt, 1982). If gamma change is detected, then this would mean that comparisons across time in work-family conflict would need to be interpreted with caution because gamma change is a threat to the validity of the interpretation that alpha change occurred (Riordan et al., 2001). When gamma change is established, the comparison of scores of the same measure across time points means something different than in the absence of gamma change (Chan, 1998; Riordan et al., 2001). If gamma change has occurred in a construct, then comparing factor means is not advised, because the qualitative shift in perception of the latent construct renders quantitative comparison unclear: if your intervention changes an apple to an orange, you should not compare them. Thus, we expect that:

Hypothesis 3: Gamma and beta changes in work-family conflict as a result of an intervention will affect the interpretation of alpha change, such that changes in the interpretation (gamma) and scaling (beta) will affect whether (and how) factor mean changes are observed.

Method

To test our hypotheses, we used field data from the publicly available Work, Family, and Health Study (WFHS). The dataset is available at https://workfamilyhealthnetwork.org/data and additional WFHS publications can be found at that site (select publications are provided in a data transparency table in Appendix A in supplemental materials). We use data from Time 1 (pre-intervention) and Time 2 (6 months post intervention) for this study, as we are concerned with the initial change effects of the intervention and not the sustainability of the intervention's effect over time (e.g., Whelan et al., 2014). Demographics from both Tomo (an information technology company consisting primarily of skilled, professional, salaried employees— most of whom had either college or advanced degrees) and Leef (a for-profit extended care organization consisting largely of hourly, less skilled employees) are provided in these studies (cf Hammer et al., 2011, 2016; Kossek, et al., 2019; Kelly et al., 2014).

Measures

Work-family conflict and family-work conflict To test our hypotheses, we use two separate measures (see Table 2). The WFHS asked participants about their work-family conflict using a well-regarded, common measure developed by Netemeyer et al., (1996). This measure has been cited by thousands of work-family scholars and assesses work-family conflict from both directions: family interfering with work and work interfering with family (bidirectional), which is supported by both theory and meta-analysis (e.g., Byron, 2005) and covers both time and strain-based conflict. Each scale consisted of five items, which were responded to on a 5-point Likert-type scale from "strongly disagree" to "strongly agree". Generally, this scale is expected to load on two correlated factors (Netemeyer et al., 1996).

Analysis: Measurement Equivalence Approach

Many features of measurement equivalence testing have analogs in organizational change assessment (see Table 1). We generally follow the steps provided by Vandenberg and Lance (2000), who prescribe a step-by-step process for equivalence testing that includes (1) configural equivalence to be sure that each group has the same factor structure, then (2) metric equivalence where the factor loadings are assumed equal across groups (using a multiple group comparison process or constraining factor loadings to equality and comparing model

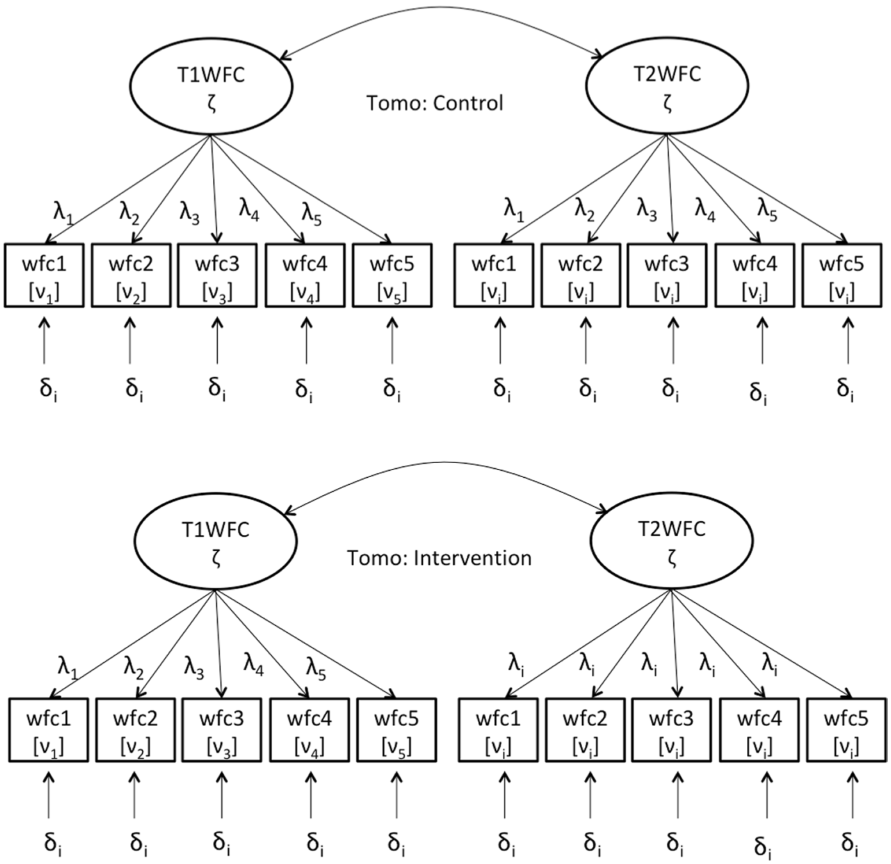
T1WFC
ζ
Tomo: Control
T2WFC
ζ
λ1 λ2 λ3 λ4 λ5
wfc1 [v1] wfc2 [v2] wfc3 [v3] wfc4 [v4] wfc5 [v5]
λ1 λ2 λ3 λ4 λ5
wfc1 [vi] wfc2 [vi] wfc3 [vi] wfc4 [vi] wfc5 [vi]
δi
T1WFC
ζ
Tomo: Intervention
T2WFC
ζ
λ1 λ2 λ3 λ4 λ5
wfc1 [v1] wfc2 [v2] wfc3 [v3] wfc4 [v4] wfc5 [v5]
λi λi λi λi λi
wfc1 [vi] wfc2 [vi] wfc3 [vi] wfc4 [vi] wfc5 [vi]
δi

Fig. 1 Demonstrating Metric Equivalence Testing Across Time and Groups, Within Organization. Equation for CFA, used in equivalence testing: $X_i = \nu_i + \lambda_i(\zeta) + \delta_i$ Where X_i = observed variable; ν_i = intercept; λ_i = factor loading; ζ = factor mean; δ_i = unique measurement error. Same subscript indicates constraints to equality. "i" subscript indicates freely estimated parameter

fit), and then (3) scalar equivalence where the intercepts are also assumed equal. For tests of measurement equivalence, the null hypothesis is that groups are equivalent; the implications of rejection of the null hypothesis depend on the specific test (Vandenberg & Lance, 2000). See Fig. 1 for an example of metric equivalence across time and intervention group (within organizations). It is only if all three steps—overall factor structure (configural), factor loadings (metric), and item intercepts (scalar)—are found to be equivalent across groups that factor means between groups can be compared (alpha change).

Testing for Gamma Change To test hypothesis 1 (gamma change), we use three methods. First, we run tests of configural equivalence (general model appropriateness demonstrating that there are two factors—WFC and FWC) by running confirmatory factor analyses for each group separately (i.e., by company, control or intervention group, at each time period) to establish whether the hypothesized factor structure fits appropriately for each group using established rules of thumb for goodness of fit (Vandenberg & Self, 1993; see Table 1). First, we compare the hypothesized two-factor model (five items loaded on a WFC factor and five items loaded on a FWC factor) to a one-factor model at each time period/in each group to make sure the model structure does not change, which would reflect a change in construct definition.

Second, we use absolute fit indices to examine overall model fit. Because these are not nested models, we should not conduct formal model comparison tests using change in X^2 or change in CFI as comparative indicators of poorer/better fit. However, the AIC (Akaike information criterion) and other absolute model fit indices can be used to compare the relative appropriateness of non-nested models by taking into account the fit and complexity of competing models. A lower AIC refers to better model fit (Burnham & Anderson, 2002). If the appropriate model structure changes over time, within the intervention group, it can be an indicator of gamma change (reconceptualization of the construct of conflict itself).

Third, we compare the equality of factor covariances across time, within company, by intervention group by examining change in X^2 and change in CFI, as per Vandenberg and Self (1993) and Vandenberg and Lance (2000) when allowing factor covariances to vary versus constraining them to equality. This represents whether the relationship between each factor (WFC and FWC) has changed over time, a reconceptualization of the constructs over time, by group. To establish whether equality constraints produce non-equivalence, we examine the chi-square differences between two nested models, but also a change in the CFI (greater than -0.002) which is interpreted as equivalent at the 0.05 type 1 error level (Meade et al., 2008). This metric is considered properly powered to detect metric and scalar non-invariance (Meade et al., 2008). As per Meade et al. (2008), we also include McDonald's Non-centrality index as a third indicator of invariance. At each stage, we report our analytical decision-making process; i.e., if these three data points (change in X^2 and change in CFI and change in McNCI) disagree, we discuss our evaluation of equivalence or non-equivalence for replication purposes. Because of the nested nature of the data, we use Satorra-Bentler X^2 values (formula included in appropriate tables; Asparouhov & Muthén, 2010; Satorra & Bentler, 2001).[2]

Testing for Beta Change To test hypothesis 2 (beta change), we use two steps of measurement equivalence testing: constraining factor loadings (metric invariance)

[2] Given the study design (employees nested in work groups), we use "Type = Complex" with clustering to utilize complex survey standard errors using the Hubert-White sandwich estimator (Muthén & Satorra, 1995) because we do not model parameters on both the between/within levels. These standard errors appropriately take into account stratification and non-independence due to clustering at the manager-level.

and item intercepts (scalar invariance) to equality across groups. Change is assessed by examining change in X^2 statistics and change in CFI to see whether there are significant differences in (1) factor loadings or (2) item intercepts. Change in factor variances is then assessed by constraining factor variances to equality over time and evaluating X^2 and change in CFI.

Then, to identify the source of beta change, we apply two methods. First, we use modification indices to examine which items are most likely to be non-equivalent (Byrne et al., 1989). Second, we use a dummy-coded covariate representing "company" or "intervention" to examine direct effects on each item using MIMIC modeling (i.e., Multiple Indicator Multiple Cause modeling or direct covariate effects; Bollen & Bauldry, 2011; Finch, 2005; Masyn, 2017). MIMIC modeling is a direct regression approach where the group variable is regressed both on the latent factors and also on the items that are equivalent to investigate whether there are direct item-level effects that may not be apparent when comparing factor means without equivalence testing (MIMIC modeling).

MIMIC modeling regresses latent constructs on covariates (grouping variables), and Muthén (1988) and Gallo et al., (1994) extended this idea to include regressing item responses on covariates directly, unmediated by the latent constructs. Direct effects indicate whether item responses differ across groups after controlling for any latent mean differences (Fleishman & Lawrence, 2003). Thus, scalar non-equivalence can be interpreted using Differential Item Functioning (DIF) terms, such that a lack of equivalence means that there are group differences in item response after accounting for the latent construct score. Item bias occurs when different item responses are caused by factors that are irrelevant to the underlying construct being measured.[3]

Testing for Alpha Change Finally, to assess hypothesis 3, we compare changes in factor means (alpha change) across two conditions: assuming invariance (e.g., holding all loadings and intercepts constant across groups) and assuming full non-invariance (e.g., letting all loadings and intercepts vary across groups). Assuming invariance means that there is an assumption of no beta or gamma change. Assuming full non-invariance means acknowledging that such other changes may exist, but not addressing them.

Sample syntax (Mplus) is provided in supporting materials.

[3] The intercept (our concern with scalar equivalence testing) is the predicted value of the measure/item when the latent factor is "0". Adding another predictor (i.e., regressing the item on an additional predictor variable) changes the intercept. Of particular note here is that the predictor added is a dichotomous variable. Thus, any significant item-level slope is the effect of the intervention ($X=1$) on the item itself, which, if we also regress the latent factor on intervention, is an effect controlling for the effect of the intervention on the latent factor itself. Davidov et al. (2015) demonstrate this approach in a cross-cultural, multilevel study; we adapt it to our comparison of two companies and of intervention and control groups.

Results

The means and correlations of the items used in our study are available in supporting materials.

As noted above, we used multiple methods to assess the presence of each type of change, as reflected by the multiple indicators of change discussed in the prior literature. To test hypothesis 1 (gamma change) we first examined configural models for each group. As shown in Table 3, the configural two-factor model reflecting FWC and WFC represented by 5 items each fit the data similarly in each group across each company sampled. We used Hu and Bentler's (1999) suggestion of RMSEA < 0.06 and SRMR < 0.08 as good fit, and the 2-index presentation strategy (to reject a model). We also compared the two-factor model shown to a one-factor model to determine whether our hypothesized factor structure fit the data best in each group at each time period.

In Table 3, we present the absolute fit statistics (AIC, BIC, X2, RMSEA, SRMR, McNCI) for each model (at both time periods, in both intervention and control groups, for both organizations, Leef and Tomo). We constrain Time 1 models in each group/organization to the same sample size as responded at Time 2 to account for changes in fit that are due to attrition and could not be associated with gamma change. Using standards for "good fit", all models met the established rules of thumb for SRMR, but only those models in Tomo met the standards for "decent" fit using RMSEA. All chi square statistics were significant, which is not unexpected given the nested nature of the data and the high sample size. First, we note that a one-factor model (all 10 items loading on a general "conflict" factor, regardless of direction) fit the data significantly worse in each group (using model comparison tests) compared to the hypothesized two-factor model. This suggests that the overall factor structure did not change as a result of the intervention. Using this indicator would suggest no gamma change is present.

Next, we examined the overall fit of each configural model, comparing the intervention group at T2 to other models within each organization. The fit in Tomo for the intervention group at T2 is the best fitting model (and is the only one to be considered a "good" fit across all three model fit statistics). The fit in Leef for the intervention group at T2 is better than the group at T1 but is not a better fitting model than any other group. Overall, this method lends some support to the hypothesis that gamma change as a result of the intervention, particularly in Tomo, may have occurred.

To probe this more, we examined the modification indices for each configural model. For the Tomo Intervention Group at Time 2, the number of cross-loaded items decreased compared to this same group at Time 1, and the magnitude of the MI for the one item that remained significant (fwc1 crossloading on the WFC factor) was greatly reduced (by > 10 points). As a note, across all groups, in both organizations, the fwc1 item was indicated as the largest, significant cross-loading item. For example, in the intervention group in Leef at time 1, the MI was 69.03 (which was reduced to 39.88 at time 2). Because the Leef absolute fit statistics were less conclusive as to changes in configural fit at Time 2 in the intervention

Table 3 Configural Equivalence Testing (Gamma Change) for all Cells using Absolute Fit Indices

	N	X-square (df)	Scaling MLR	RMSEA	SRMR	BIC	AIC	McNCI	Cross-loaded items
Leef (intervention, T1)	585	181.14 (34)	1.297	.086	.066	13,312.14	13,447.61	.882	4
Leef (intervention, T2)	585	170.19 (34)	1.346	.083	.062	12,674.90	12,539.38	.890	3
Leef (control, T1)	670	153.23 (34)	1.363	.072	.053	14,262.95	14,402.68	.915	3
Leef (control, T2)	670	162.74 (34)	1.467	.075	.051	13,409.83	13,270.11	.908	3
Tomo (intervention, T1)	353	101.82 (34)	1.227	.075	.064	7729.19	7609.32	.908	3
Tomo (intervention, T2)	353	81.22 (34)	1.281	.063	.049	7024.50	6904.64	.935	1
Tomo (control, T1)	325	90.55 (34)	1.172	.072	.069	7152.06	7269.36	.916	3
Tomo (control, T2)	325	98.96 (34)	1.309	.077	.066	6927.81	6810.51	.905	4

Fit statistics are reported to three decimal places throughout to facilitate comparisons to rules of thumb. Time 1 sample size is constrained based on attrition at time 2 to facilitate comparison (replicating listwise deletion). McDonald's Relative NonCentrality Index (McNCI) > .90 is considered "good fit" (Hu & Bentler, 1999). Sample size reflects adjustments for listwise deletion across time. Only participants who responded at both T1 and T2 were included group compared to other groups, we cannot conclude that the evidence is consistent for intervention caused gamma change in Leef. With that said, this method demonstrated the possibility for gamma change in model configuration as a result of the intervention in Tomo (providing some support for H1).

As a final test of gamma change, we examined the consistency of factor covariances at each time period. The two factors measured at each time period were WFC and FWC, and, if gamma change were observed, the covariance between these two factors may differ over time in the intervention group as compared to the control group, within each company, indicating a change in structure. Constraining the factor covariances between WFC and FWC to equality in the intervention and control groups at Times 1 andn 2 did not result in worse fit for Tomo ($\Delta X^2[1]=2.32$; $\Delta CFI=0.000$) or Leef ($\Delta X^2[1]=0.01$; $\Delta CFI=0.001$). The constrained covariance between WFC and FWC at each time period for Tomo was 0.203 and for Leef was 0.212. This does not provide support for H1 using this method.

Overall, the intervention did not seem to lead to gamma change in Tomo or in Leef in terms of overall factor structure or factor covariances—or how the work-to-family conflict construct and the family-to-work construct are perceived overall or in relation to each other, but the intervention may have shifted the interpretation of the construct somewhat by reducing cross-loading items and making the constructs clearer, particularly in the Tomo organization. This provides some support for H1 in Tomo.

To assess beta changes (hypothesis 2), we ran metric and scalar invariance tests. If beta change had occurred due to the intervention, we would expect less invariance/equivalence in factor loadings and intercepts in the intervention groups at Time 2 (post-intervention), compared to Time 1 and the control group at both time points. In Table 4, the "configural invariance across groups" row is the factor model where no metric or scalar constraints are included, but both control/intervention and Time 1/Time 2 factors are included in the same model for each company separately. This is the model that assumes the configural invariance that we tested in Table 3. In Table 4, we present multiple comparative fit statistics to evaluate whether the fit in a model with additional constraints on equality was equivalent to one with fewer degrees of freedom, following our approach detailed in our analysis section. Given the large sample size, and how this may affect interpretation of X^2 statistics, we determined that if three metrics indicate inequivalence, we have interpreted the fit as being non-equivalent (shown in the final column in Table 4).

As shown in Table 4, we achieve metric equivalence across groups and across Time 1 and 2 in both companies. Using change in X-square, change in McNCI and change in CFI as indicators of non-invariance, there was scalar non-invariance in Tomo for the control group at T2. This suggests some shifting in scaling of the constructs (beta change) that is just due to overall changes over time (perhaps contamination from the intervention affecting non-treatment groups) or general shifts over time. Because this is not in the intervention group, it is not intervention-related change. After we made adjustments to two items (freeing the intercepts of wfc5 and wfc1 from their equality constraints), we achieved invariance and moved on to the intervention group at T2. There was metric invariance (as shown in Table 4), but not scalar invariance. The scalar non-invariance at T2

Table 4 Hypothesis 2: Beta Changes in WFC and FWC

Model	N	C_0	df	X^2	CFI	Mc	ΔCFI	ΔMc	Satorra-Bentler ΔX^2	Δdf	p-value	Interpretation
TOMO												
Configural across groups	781	1.190	328	790.120	.938	.744						
Metric T1		1.189	336	801.117	.938	.742	.000	-.001	10.70	8	.219	Equivalent
Scalar T1		1.190	346	805.739	.939	.745	.001	+.003	5.15	10	.881	Equivalent
Metric post (control)		1.189	354	815.632	.938	.744	-.001	-.001	9.56	8	.297	Equivalent
Scalar post (control)		**1.182**	**364**	**851.848**	**.935**	**.731**	**-.002**	**-.012**	**39.71**	**10**	**<.001**	**Not Equivalent**
Freed intercepts at Time 2 (wfc5; wfc1)		1.183	362	835.151	.937	.738	-.001	-.005	19.83	8	.01	Equivalent
Metric post (intervention)		1.182	370	845.175	.937	.737	.001	-.001	9.68	8	.28	Equivalent
Scalar post (intervention)		**1.180**	**380**	**867.855**	**.935**	**.731**	**-.003**	**-.006**	**22.69**	**10**	**.01**	**Not Equivalent**
Freed intercepts at Time 2 (fwc2; wfc5)		1.180	378	858.212	.936	.735	-.001	-.002	12.59	8	.13	Equivalent
LEEF												
Configural across groups	1502	1.258	328	1232.83	.916	0.740						
Metric T1		1.257	336	1236.55	.916	0.741	.000	-.001	2.83	8	.944	Equivalent
Scalar T1		1.253	346	1261.10	.915	0.737	-.001	-.004	23.08	10	.010	Equivalent
Metric post (control)		1.252	354	1272.62	.915	0.736	.000	-.001	10.90	8	.207	Equivalent
Scalar post (control)		1.245	364	1290.20	.914	0.735	-.001	-.001	13.00	10	.224	Equivalent
Metric post (intervention)		1.249	372	1288.36	.915	0.737	+.001	+.002	2.00	8	.981	Equivalent
Scalar post (intervention		1.242	382	1306.70	.914	0.735	-.001	-.002	14.02	10	.172	Equivalent

Bolded rows indicate non-equivalent model fit. Satorra-Bentler X2 = (F0*c0- F1*c1)(d0- d1)/(c0*d0- c1*d1) Where F1 = X^2 of the less restrictive model; F0 = X^2 of the more restrictive model; d1 = degrees of freedom of the less restrictive model; d0 = degrees of freedom of the more restrictive model; c1 = correction factor of the less restrictive model; c0 = correction factor of the more restrictive model. Mc: McDonald's Noncentrality Parameter. AFI cutoffs from Meade et al. (2008): ΔCFI < .002; ΔMc (for 5 items per four factors constrained to equality) < .0058. Lack of invariance = 3 of 3 (significant ΔCFI, ΔMc, significant ΔX^2.) Degrees of freedom are the same for each organization (Leef and Tomo) because the items and models (thus the parameters estimated—including factors, covariances, and variances) are identical. Sample size reflects Mplus default missing data approaches, using maximum likelihood and missing at random (MAR)

Table 5 Intercepts for Tomo Denoting Scalar Invariance and Non-Invariance

Item	Intervention	Control	Item	Intervention	Control
T1 wfc1	3.106	3.106	T2 wfc1	3.106	**3.000**
T1 wfc2	2.803	2.803	T2 wfc2	2.803	2.803
T1 wfc3	2.917	2.917	T2 wfc3	2.917	2.917
T1 wfc4	2.848	2.848	T2 wfc4	2.848	2.848
T1 wfc5	3.316	3.316	T2 wfc5	**3.200**	**3.148**
T1 fwc1	2.444	2.444	T2 fwc1	2.444	2.444
T1 fwc2	2.142	2.142	T2 fwc2	**2.226**	2.142
T1 fwc3	1.986	1.986	T2 fwc3	1.986	1.986
T1 fwc4	2.041	2.041	T2 fwc4	2.041	2.041
T1 fwc5	1.986	1.986	T2 fwc5	1.986	1.986

Bold font in the table represents item intercepts which are not-invariant across groups at Time 2

in the intervention group does provide some evidence supporting H2—that the intervention was associated with beta change related to item intercepts. Examining modification indices identified the intercepts for items fwc2 and wfc5. Freeing these intercepts to vary across time established invariance.

In Table 5, we present all of the intercepts in Tomo and indicate using bold font where the non-invariant intercepts were identified. The magnitude of the shift does not seem large, (e.g., 2.041 for fwc2 to 2.226 at T2 in the intervention group), but it represents a quarter of a standard deviation shift for that item that is not related to the latent construct of FWC itself, but represents the intercept of the item and a mean shift in that item score irrespective of the construct of FWC. Overall, this provides some support for H2 in Tomo—that the intervention was associated with scalar-related non-invariance, particularly in the intercepts for two items.

Using MIMIC modeling to address the degree to which this scalar non-invariance may be indicative of beta change in hypothesis 2, in which, instead of using multiple group comparison, we regress each item onto a dummy variable for "intervention vs. control". Controlling for alpha change (regressing the factor means on the intervention dummy), the intervention had a *positive* effect on fwc2's intercept (B = 0.08, p = 0.025) but no direct effect on wfc5's intercept (B = -0.06, p = 0.31). We also demonstrate how a (potentially unintended) shift in interpretation of the scale itself (in which responses on the item "Put off things at work because of demands on time at home" are higher at Time 2) unrelated to the amount of FWC experienced (the latent factor) may also obscure evaluations of alpha change (the intervention effect on FWC at T2, controlling for FWC at T1 was 0.01, p = 82; the intervention effect on WFC at T2, controlling for T1, was -0.08, p = 0.06 using this method of assessing alpha change). This confirms our partial support for H2 in Tomo.

The final check of beta change involved examining the equivalence of factor variances across time. Thus, we checked for factor non-invariance using the same measurement invariance approach. Constraining the variances to equality across time and group for each scale (WFC and FWC) indicated no decrease in fit using ΔCFI

Table 6 Hypothesis 3: The Effect of Equivalence Testing on Alpha Change

LEEF								
	Assuming No Equivalence (all parameters freely estimated)				Assuming Full Equivalence (all parameters constrained to equality)			
	Mean	SD	t-statistic	p-value	Mean	SD	t-statistic	p-value
WFC2	-.08	.05	-1.43	.15	-.04	.03	-1.13	.26
FWC2	**-.08**	**.04**	**-1.81**	**.07**	**.01**	**.02**	**.43**	**.67**
TOMO								
	Assuming No Equivalence (all parameters freely estimated)				Assuming Full Equivalence (all parameters constrained to equality)			
	Mean	SD	t-statistic	p-value	Mean	SD	t-statistic	p-value
WFC2	**.11**	**.07**	**1.54**	**.12**	**-.08**	**.04**	**-1.90**	**.057**
FWC2	.03	.06	.46	.65	.03	.04	.92	.36

Bolded rows indicate non-equivalent model fit. Intervention group factor means are reported; control group factor means and time 1 factor means for both groups are constrained to zero. P-values < .05 represent significant differences in factor means for these outcomes versus the control group (and Time 1). With freely estimated models to facilitate mean comparison, model identification is achieved via constraining one factor loading and one intercept to equality

(0.000 for Leef, + 0.001 for Tomo) or ΔX^2 (not significant) or ΔMcDonald's NCI (0.001 for both). This does not support H2 for factor variances.

Finally, to test hypothesis 3, we determined whether changes in factor means in the intervention group across time (evidence of alpha change) seemed to be affected by whether or not potential gamma or beta change is accounted for in the model. In Table 6, we start with a baseline analysis that ignores potential beta or gamma change, allowing for freely estimated factor loadings and intercepts, and covariances and variances across group/time.[4] We compare this analysis to one that assumes full metric and scalar equivalence (and factor covariances and variances, per the organizational change literature, see Table 1). We then constrained the factor means in the control group to zero, and the factor means at Time 1 in the intervention group to 0 as well. This makes the factor mean at Time 2 in the intervention group an effective test of alpha change, i.e., the change in the factor mean that can be attributed to the intervention. As shown in Table 6, assumptions of no beta/gamma change (e.g., invariance) affect the interpretation of alpha change compared to models where such change is not controlled for.

Thus, our conclusions, even the most basic of tests of the intervention's effect on levels of work-family conflict (e.g., assessing alpha change using factor mean

[4] Comparing to Kelly et al. (2014) which used a multilevel analysis (employees nested in study groups) as we do here, looking only at time 2 (the immediate survey after intervention) and the interaction of time 2 and intervention status as the intervention effect (including random slopes). Since our purpose is not to pinpoint the precise effect of the intervention (which the aforementioned authors already did, finding a significant effect of intervention on time 2 FWC but not WFC in Tomo), but rather to compare effects given different organizational change/measurement equivalence assumptions, we do not replicate their analysis exactly.

comparisons across group and time) might be affected by ignoring beta and gamma change. This provides support for hypothesis 3.

Discussion

Research designs in the occupational health intervention literature are becoming increasingly sophisticated—including repeated-measures, interventions, and field experiments, which is particularly reflected in work-family interventions like the STAR intervention we assess here (e.g., Hammer et al., 2011). We demonstrate how improved integration of organizational change and measurement equivalence literatures is important to advance the occupational health science of work-family intervention research. We do so using evidence from an established randomized controlled trial, the gold-standard for studying organizational change. Kelly et al. (2014), used a nested approach to test the effect of the intervention in Tomo at each post-intervention time period, using individual-level scale scores, as opposed to latent factors, finding that there was a negative effect of the intervention on time 2 FWC (alpha change). In our research, we expand this test to include tests of the intervention on both gamma and beta types of organizational change using measurement equivalence analyses.

We find results consistent with some gamma and beta change as a result of the intervention in Tomo, but not in Leef. These other types of change were not assessed in Kelly et al's (2014) assessment that was focused on alpha change. Kelly et al. (2014) did not expect gamma or beta change, thus it makes sense that they did not test for it. After all, as Riordan et al. (2001) note, few (if any) tests of intervention efficacy do so. But our findings that the STAR intervention may have affected change in different ways than just via overall work-family and family-work conflict scores suggests a missed opportunity for evaluators. Our findings support the need for studies to more closely integrate measurement equivalence methodology considerations into intervention research when considering change. Our paper provides a new approach to help advance understanding of assessing various types of organizational change in response to work-family interventions.

Implications for Research and Practice

The alpha, beta, gamma change approach has yet to be employed as a theoretical model in response to organiazational interventions. The measurement invariance approach, combined with direct effects assessed using MIMIC modeling, can directly assess hypotheses related to these different types of change. Thinking about interventions with a focus on how the intervention is expected to affect change in individuals and organizations can help improve the fidelity between such theorizing and the assessments used to evlaaute intervention effectiveness. Is the intervention meant to affect how people define a construct (e.g., gamma change), such as a sexual harassment education training? Perhaps looking at factor structures and cross-loading items can help to assess whether this intervention was effective. Is the intervention meant to shift how people see stressful situations and shift how they respond to items about their stress (e.g., beta change)? Perhaps looking at item intercepts is a more direct assessment of its effectiveness. This may be particularly important for outcomes of interest that are measured using measure that do not hang together as a latent construct as well (e.g., "formative" measures; Diamantopoulos & Winklhofer, 2001) as the individual items can be disaggregated in ways that pinpoint functions of change induced by the intervention.

Measures such as work-family conflict that are commonly used in the work family field are an example of such measures. We believe that change in multiple outcomes of interest that are less susceptible to beta change such as archival measures of actual use of greater flexibility, or turnover measures (perhaps including people who have actually left the organization since the intervention began) are examples of hard measures that are needed to augment formative measures that are susceptible to gamma change. Using both in conjunction with one another could also help researchers to understand the intervention's mechanism of change even more clearly, as we found that multiple types of change in response to an intervention could co-exist.

Though researchers and practitioners are constantly searching for generalizable evidence-based work-family interventions that can be replicated across contexts and demands, these assumptions should be tested empirically—and not merely via factor mean comparisons. Research on national cultures require the establishment of measurement equivalence (Ryan et al., 1999), but we rarely see this reported in comparisons of intervention efficacy. Our paper demonstrates that measurement non-equivalence can and should be interpreted, understood, and even hypothesized as indicative of planned change.

Using measurement invariance/equivalence approaches can help assess different types of organizational change in response to an intervention, but can also be used to assess change or that is a result of other causes. For example, in this study, there was also scalar non-invariance in Tomo for the control group across time, showing that there were non-intervention-related beta changes that we demonstrated using this methodology. This change may have been related to non-study-related issues (e.g., see Lam et al., 2015 for a merger in this company that occurred during that time) or due to study-related contamination between the control and the intervention group.

Respondents in the control and intervention sites may respond to work-family measures in different ways due to their social contexts, the unique culture or climate of their particular work environment, or to norms about supportive behavior in their groups (Thompson et al., 1999). This is particularly true when cluster randomized designs are used. For instance, respondents may interpret scales differently, maybe perceiving "demands of the job", or "duties at home" differently (Netemeyer et al., 1996), over time, or even perceive their organizational culture differently based on their contexts, which have nothing to do with the intervention itself. This suggests that it is particularly important to pay attention to norm shifts, as these domains are permeated with strong societal norms and expectations. It is also possible that these norm shifts over time could affect the effect of the intervention, by making an intervention more or less effective, depending on whether we controlled for other mechanisms of change. This could provide important feedback to practitioners looking to roll out intervention pilots to broader contexts where there is more variance in contexts. While control variables and random assignment can help in these cases, possible contamination across groups or changes not captured by control variables are still important to capture.

Future research opportunities (and practical considerations) that expand on traditional conceptualizations of organizational change in this context abound, such as expecting whether people in similar demographic groups will report more or less beta or gamma change due to interventions, due to shared norms and realities, that can be examined. For instance, say an intervention to provide more work schedule control was instituted in an organization. We might expect that the effect of such intervention may come out in gamma change (interpreting the construct of "control" differently) or beta change (rethinking what "high" control versus "low" control is). But these beta and gamma changes may be moderated by demographic groups—where the differential effects of such an intervention may come out via the assessment of change in different ways. Or, for instance, older and younger women may perceive a measure of "caregiving" differently based on their past experiences with caring for children or aging parents, which may be important to control for prior to asssessing the alpha change effects of a caregiving support intervention. Given the COVID-19 pandemic homeworkers may view constructs assessing satisfaction, teleworking and place flexibility differently since it was a forced shift. These and many more questions can be substantively addressed using this approach.

While the results of this study do not necessarily imply that the occupational health psychology field needs a wholesale reassessment of intervention studies, we do believe that in conjunction with general checklists for measurement equivalence (Van de Schoot et al., 2012), our paper can serve as a useful guide to work-family researchers of how the processes involved in testing for measurement equivalence can be used to assess different types of intervention-related change. Specifically, intervention researchers should consider carefully what their theory of change is for their intervention (see Reinholz & Andrews, 2020; De Silva et al., 2014) and make a priori assumptions about how the theory will affect their constructs of interest. They should select measures of constructs that will allow them to assess such change. And they should use measurement equivalence approaches where appropriate to assess whether this sort of change has occurred. Finally, they should remember that an accounting for gamma and beta change is necessary prior to assessing alpha change, given that change not-related to the intervention could also be present when evaluating measures in groups over time (as we see in the Tomo control group in this study). Overall, we think that our results provide many opportunities for interventionists to think more clearly and carefully about how and why their interventions will create change, and assess them accordingly.

Conclusion

The efficacy of interventions designed to improve occupational health and how they create organizational change are critical questions for researchers and practitioners alike. Our research supports the view that change from interventions can occur through multiple mechanisms which are often perceived as mostly methodological artifacts of measurement, and that this should be considered, and even theorized, when conceptualizing intervention studies. As interventions, longitudinal, and experience sampling methodologies grow more common in occupational health studies, understanding different mechanisms of change using a measurement equivalence approach can be a useful tool for researchers to create replicable and generalizable research and for practitioners who seek to apply our findings to their workplaces.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41542-022-00122-y.

Authors' Contributions All authors contributed to the paper conception. Ellen Kossek, Rebecca Thompson, and Todd Bodner participated in original data collection and preparation in association with the funding and publicly available data presented above. The first draft of the manuscript was written by Beth Livingston and Shaun Pichler, with data analysis by Beth Livingston. Revisions were led by Beth Livingston and Shaun Pichler. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The publicly available data used in this study was supported by the United States Department of Health and Human Services. National Institutes of Health. Eunice Kennedy Shriver National Institute of Child Health and Human Development (U01HD051217, U01HD051218), United States Department of Health and Human Services. National Institutes of Health. National Institute on Aging (U01AG027669), United States Department of Health and Human Services. Centers for Disease Control and Prevention. National Institute of Occupational Safety and Health (U01OH008788, U01HD059773), United States Department of Health and Human Services. National Institutes of Health. National Heart Lung and Blood Institute (R01HL107240), William T. Grant Foundation, Alfred P. Sloan Foundation, United States Department of Health and Human Services. Administration for Children and Families.

Data Availability The data that support the findings of this study are openly available at the Work, Family and Health Network. Work, Family, and Health Study (WFHS) Inter-university Consortium for Political and Social Research [distributor], 2018–10-03. https://doi.org/10.3886/ICPSR36158.v2

Code Availability Code is presented in supplemental materials.

Declarations

Additional declarations for articles in life science journals that report the results of studies involving humans and/or animals Not applicable

Ethics Approval (include appropriate approvals or waivers) Publicly available data was used.

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Allen, T. D., & Martin, A. (2017). The work-family interface: A retrospective look at 20 years of research in JOHP. *Journal of Occupational Health Psychology, 22*(3), 259–272.

Asparouhov, T., & Muthén, B. (2010). Computing the strictly positive Satorra-Bentler chi-square test in Mplus. *Mplus Web Notes*, 12. Retrieved from: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.310.3465&rep=rep1&type=pdf

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly*, *36*, 421–458.

Barber, L. K., & Santuzzi, A. M. (2015). Please respond ASAP: Workplace telepressure and employee recovery. *Journal of Occupational Health Psychology, 20*(2), 172–189.

Bennett, M. M., Beehr, T. A., & Ivanitskaya, L. V. (2017). Work-family conflict: Differences across generations and life cycles. *Journal of Managerial Psychology, 32*(4), 314–332.

Bhave, D. P., Kramer, A., & Glomb, T. M. (2010). Work–family conflict in work groups: Social information processing, support, and demographic dissimilarity. *Journal of Applied Psychology, 95*(1), 145–158.

Bodner, T. E., & Bliese, P. D. (2018). Detecting and differentiating the direction of change and intervention effects in randomized trials. *Journal of Applied Psychology, 103*(1), 37–53.

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods, 16*(3), 265–284.

Breuer, E., Lee, L., De Silva, M., & Lund, C. (2015). Using theory of change to design and evaluate public health interventions: A systematic review. *Implementation Science, 11*(1), 1–17.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer-Verlag.

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for equivalence of factor covariance and mean structures: The issue of partial measurement equivalence. *Psychological Bulletin, 105*, 456–466.

Byron, K. (2005). A meta-analytic review of work–family conflict and its antecedents. *Journal of Vocational Behavior, 67*(2), 169–198.

Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organizational Research Methods, 1*(4), 421–483.

Chan, D., & Schmitt, N. (2000). Interindividual differences in intraindividual changes in proactivity during organizational entry: A latent growth modeling approach to understanding newcomer adaptation. *Journal of Applied Psychology, 85*(2), 190–210.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005–1018.

Cheung, G. W., & Lau, R. S. (2012). A direct comparison approach for testing measurement equivalence. *Organizational Research Methods, 15*(2), 167–198.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255.

Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology, 55*(4), 584.

Davidov, E., Cieciuch, J., Meuleman, B., Schmidt, P., Algesheimer, R., & Hausherr, M. (2015). The comparability of measurements of attitudes toward immigration in the European Social Survey: Exact versus approximate measurement equivalence. *Public Opinion Quarterly, 79*(S1), 244–266.

De Boer, A. G. E. M., Van Beek, J. C., Durinck, J., Verbeek, J. H. A. M., & Van Dijk, F. J. H. (2004). An occupational health intervention programme for workers at risk for early retirement; a randomised controlled trial. *Occupational and Environmental Medicine, 61*(11), 924–929.

De Silva, M. J., Breuer, E., Lee, L., Asher, L., Chowdhary, N., Lund, C., & Patel, V. (2014). Theory of change: A theory-driven approach to enhance the Medical Research Council's framework for complex interventions. *Trials, 15*(1), 1–13.

Diamantopoulos, A. (2011). Incorporating formative measures into covariance-based structural equation models. *MIS Quarterly, 35*, 335–358.

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research, 38*(2), 269–277.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278–295.

Fleishman, J. A., & Lawrence, W. F. (2003). Demographic variation in SF-12 scores: true differences or differential item functioning?. *Medical Care*, *41*, III75–III86.

Gallo, J. J., Anthony, J. C., & Muthén, B. O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journal of Gerontology, 49*(6), 251–264.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science, 12*(2), 133–157.

Gorin, S. S., Badr, H., Krebs, P., & Das, I. P. (2012). Multilevel interventions and racial/ethnic health disparities. *Journal of the National Cancer Institute Monographs, 2012*(44), 100–111.

Hammer, L. B., Johnson, R. C., Crain, T. L., Bodner, T., Kossek, E. E., Davis, K. D. … & Berkman, L. (2016). Intervention effects on safety compliance and citizenship behaviors: Evidence from the work, family, and health study. *Journal of Applied Psychology, 101*(2), 190-208.

Hammer, L. B., Kossek, E. E., Anger, W. K., Bodner, T., & Zimmerman, K. L. (2011). Clarifying work–family intervention processes: The roles of work–family conflict and family-supportive supervisor behaviors. *Journal of Applied Psychology, 96*(1), 134–150.

Hammer, L. B., Kossek, E. E., Bodner, T., & Crain, T. (2013). Measurement development and validation of the family supportive supervision behavior short-form (FSSB-SF). *Journal of Occupational Health Psychology, 18*, 285–296.

Hammer, L. B., Kossek, E. E., Yragui, N. L., Bodner, T. E., & Hanson, G. C. (2009). Development and validation of a multidimensional measure of family supportive supervisor behaviors (FSSB). *Journal of Management, 35*(4), 837–856.

Hofmans, J., Dries, N., & Pepermans, R. (2008). The Career Satisfaction Scale: Response bias among men and women. *Journal of Vocational Behavior, 73*(3), 397–403.

Holman, D., & Axtell, C. (2016). Can job redesign interventions influence a broad range of employee outcomes by changing multiple job characteristics? A quasi-experimental study. *Journal of Occupational Health Psychology, 21*(3), 284–295.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.

Huang, S. L., Li, R. H., Huang, F. Y., & Tang, F. C. (2015). The potential for mindfulness-based intervention in workplace mental health promotion: Results of a randomized controlled trial. *PLoS One, 10*(9), e0138089.

Judge, T. A., Ilies, R., & Scott, B. A. (2006). Work–family conflict and emotions: Effects at work and at home. *Personnel Psychology, 59*(4), 779–814.

Kahn, R. L., Wolfe, D. M., Quinn, R. P., Snoek, J. D., & Rosenthal, R. A. (1964). *Organizational stress: Studies in role conflict and ambiguity.* John Wiley.

Keefer, K. V., Holden, R. R., & Parker, J. D. (2013). Longitudinal assessment of trait emotional intelligence: Measurement invariance and construct continuity from late childhood to adolescence. *Psychological Assessment, 25*(4), 1255–1272.

Kelly, E. L., Moen, P., Oakes, J. M., Fan, W., Okechukwu, C., Davis, K. D., … & Mierzwa, F. (2014). Changing work and work-family conflict: Evidence from the work, family, and health network. *American Sociological Review, 79*(3), 485-516.

Kelly, E. L., Moen, P., & Tranby, E. (2011). Changing workplaces to reduce work-family conflict: Schedule control in a white-collar organization. *American Sociological Review, 76*(2), 265–290.

Kossek, E. (2016). Implementing organizational work-life interventions: Toward a triple bottom line. *Community Work and Family*, *19* (2), 242–256. https://doi.org/10.1080/13668803.2016.1135540

Kossek, E., Hammer, L., Kelly, E., & Moen, P. (2014). Designing organizational work, family & health change Initiatives. *Organizational Dynamics, 43*, 53–63.

Kossek, E., & Lautsch, B. (2018). Work-life flexibility for whom? Occupational status and work-life inequality in upper, middle, and lower level jobs. *Academy of Management Annals, 12*(1), 5–36.

Kossek, E. E., Thompson, R. J., Lawson, K. M., Bodner, T., Perrigino, M., Hammer, L. B., Buxton, O. M., Almeida, D. M., Moen, P., Hurtado, D., Wipfli, B., Berkman, L. F., & Bray, J. W. (2019). Caring for the elderly at work and home: Can a randomized organizational intervention improve psychological health? *Journal of Occupational Health Psychology, 24*(1), 36–54. https://doi.org/10.1037/ocp0000104

Kossek, E., Wipfli, B., Thompson, R., Brockwood, K. & the Work Family Health Network Writing Team (2017). The Work, Family & Health Network intervention: Core elements and customization for diverse occupational health contexts, In Leong, F., Eggerth, D., Chang, D. Flynn, M., Ford, K. & Martinez, R. *Occupational Health Disparities among Racial and Ethnic Minorities: Improving the Well-being of Racial and Ethnic Minorities* (pp. 181–215). APA.

Lam, J., Fox, K., Fan, W., Moen, P., Kelly, E., Hammer, L., & Kossek, E. (2015). Manager characteristics and employee job insecurity around a merger announcement: The role of status and crossover. *Sociology Quarterly, 56*, 558–580.

Logan, M. S., & Ganster, D. C. (2005). An experimental evaluation of a control intervention to alleviate job-related stress. *Journal of Management, 31*(1), 90–107.

Marsh, H. W., & Grayson, D. (1994). Longitudinal stability of latent means and individual differences: A unified approach. *Structural Equation Modeling: A Multidisciplinary Journal, 1*(4), 317–359.

Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(2), 180–197.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568–592.

Menard, S. (1991). *Longitudinal research*. Newbury Park: Sage.

Millsap, R. E., & Hartog, S. B. (1988). Alpha, beta, and gamma change in evaluation research: A structural equation approach. *Journal of Applied Psychology, 73*(3), 574–584.

Moen, P., Kelly, E. L., Fan, W., Lee, S. R., Almeida, D., Kossek, E. E., & Buxton, O. M. (2016). Does a flexibility/support organizational initiative improve high-tech employees' well-being? Evidence from the work, family, and health network. *American Sociological Review, 81*(1), 134–164.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*(4), 557–585.

Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 213–238). Educational Testing Service.

Muthén, L. K., & Muthén, B. (2015). *Mplus. The comprehensive modelling program for applied researchers: User's guide.* Los Angeles, CA.

Muthén, B. O., & Satorra, A. (1995). Technical aspects of Muthén's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika, 60*(4), 489–503.

Netemeyer, R. G., Boles, J. S., & McMurrian, R. (1996). Development and validation of work–family conflict and family–work conflict scales. *Journal of Applied Psychology, 81*(4), 400.

Nielsen, K., & Randall, R. (2012). The importance of employee participation and perceptions of changes in procedures in a teamworking intervention. *Work & Stress, 26*(2), 91–111.

Nye, C. D., Newman, D. A., & Joseph, D. L. (2010). Never say "always"? Extreme item wording effects on scalar equivalence and item response curves. *Organizational Research Methods, 13*(4), 806–830.

Odle-Dusseau, H. N., Hammer, L. B., Crain, T. L., & Bodner, T. E. (2016). The influence of family-supportive supervisor training on employee job performance and attitudes: An organizational work–family intervention. *Journal of Occupational Health Psychology, 21*(3), 296–308.

Oreg, S., Bayazıt, M., Vakola, M., Arciniega, L., Armenakis, A., Barkauskiene, R., … & Hřebíčková, M. (2012). Measurement equivalence of the dispositional resistance to change scale. In *Cross-Cultural Analysis: Methods and Applications* (pp. 249–278). Taylor and Francis.

Pejtersen, J. H., & Kristensen, T. S. (2009). The development of the psychosocial work environment in Denmark from 1997 to 2005. *Scandinavian Journal of Work, Environment & Health, 35*, 284–293.

Pitts, S. C., West, S. G., & Tein, J. Y. (1996). Longitudinal measurement models in evaluation research: Examining stability and change. *Evaluation and Program Planning, 19*(4), 333–350.

Reinholz, D. L., & Andrews, T. C. (2020). Change theory and theory of change: What's the difference anyway? *International Journal of STEM Education, 7*(1), 1–12.

Riordan, C. M., Richardson, H. A., Schaffer, B. S., & Vandenberg, R. J. (2001). Alpha, beta, and gamma change: A review of past research with recommendations for new directions. In Eds. Chester A. Schriesheim, Linda L. Neider, *Equivalence in Measurement* (pp. 51–97). Information Age Publishing: CT.

Ryan, A. M., Chan, D., Ployhart, R. E., & Slade, L. A. (1999). Employee attitude surveys in a multinational organization: Considering language and culture in assessing measurement equivalence. *Personnel Psychology, 52*(1), 37–58.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507–514.

Schaubroeck, J., & Green, S. G. (1989). Confirmatory factor analytic procedures for assessing change during organizational entry. *Journal of Applied Psychology, 74*(6), 892–900.

Schmitt, N. (1982). The use of analysis of covariance structures to assess beta and gamma change. *Multivariate Behavioral Research, 17*(3), 343–358.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*(4), 210–222.

Schwartz, C. E., Sprangers, M. A., Carey, A., & Reed, G. (2004). Exploring response shift in longitudinal data. *Psychology & Health, 19*(1), 51–69.

Spurk, D., Abele, A. E., & Volmer, J. (2011). The career satisfaction scale: Longitudinal measurement invariance and latent growth analysis. *Journal of Occupational and Organizational Psychology, 84*(2), 315–326.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292–1306.

Taimela, S., Malmivaara, A., Justen, S., Läärä, E., Sintonen, H., Tiekso, J., & Aro, T. (2008). The effectiveness of two occupational health intervention programmes in reducing sickness absence among employees at risk. Two randomised controlled trials. *Occupational and Environmental Medicine, 65*(4), 236–241.

Thompson, C. A., Beauvais, L. L., & Lyness, K. S. (1999). When work–family benefits are not enough: The influence of work–family culture on benefit utilization, organizational attachment, and work–family conflict. *Journal of Vocational Behavior, 54*(3), 392–415.

Trickett, E. J., & Beehler, S. (2013). The ecology of multilevel interventions to reduce social inequalities in health. *American Behavioral Scientist, 57*(8), 1227–1246.

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement equivalence. *European Journal of Developmental Psychology, 9*(4), 486–492.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement equivalence methods and procedures. *Organizational Research Methods, 5*(2), 139–158.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement equivalence literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70.

Vandenberg, R. J., & Self, R. M. (1993). Assessing newcomers' changing commitments to the organization during the first 6 months of work. *Journal of Applied Psychology, 78*(4), 557–568.

Whelan, J., Love, P., Pettman, T., Doyle, J., Booth, S., Smith, E., & Waters, E. (2014). Cochrane update: Predicting sustainability of intervention effects in public health evidence: Identifying key elements to provide guidance. *Journal of Public Health, 36*(2), 347–351.

Work Family Health Study. (2018). Data Sharing for Demographic Research https://www.icpsr.umich.edu/icpsrweb/DSDR/studies/36158; https://doi.org/10.3886/ICPSR36158.v2 ; Retrieved June 12, 2019.

Zand, D. E., & Sorensen, R. E. (1975). Theory of change and the effective use of management science. *Administrative Science Quarterly*, *20*, 532–545.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Beth A. Livingston[1] · Shaun Pichler[2] · Ellen Ernst Kossek[3] · Rebecca J. Thompson[4] · Todd Bodner[5]

✉ Beth A. Livingston
Beth-livingston@uiowa.edu

Shaun Pichler
shaunpichler@gmail.com

Ellen Ernst Kossek
ekossek@purdue.edu

Rebecca J. Thompson
rebeccajthompson85@gmail.com

Todd Bodner
tbodner@pdx.edu

[1] Tippie College of Business, University of Iowa, Iowa City, IA W276 PBB, USA

[2] California State University, Fullerton, CA, USA

[3] Purdue University, West Lafayette, IN, USA

[4] ICF International, Fairfax, VA, USA

[5] Portland State University, Portland, OR, USA