Running Head. Multilevel Assessment and Prediction

Doubly Latent Multilevel Procedures for Organizational Assessment and Prediction

Alexandre J.S. Morin*
Substantive-Methodological Synergy Research Laboratory, Concordia University
Ann-Renée Blais*
Statistics Canada
& Department of National Defence, Canada
Léandre-Alexis Chénard Poirier
Département de psychologie, Université du Québec à Montréal

* The first two authors (A.J.S.M. & A.-R.B.) contributed equally to this article and their order was determined at random: Both should thus be considered first authors.

Acknowledgements: Preparation of this paper was supported by grants from the Canadian Institute for Military and Veteran Health Research (CIMVHR) and from the Social Science and Humanity Research Council of Canada (435-2018-0368).

Corresponding author:
Alexandre J.S. Morin, Substantive-Methodological Synergy Research Laboratory
Department of Psychology, Concordia University
7141 Sherbrooke W, Montreal, QC, Canada, H3B 1R6
Email: alexandre.morin@concordia.ca

This is the final prepublication version of:
Morin, A.J.S., Blais, A.-R., & Chénard-Poirier, L.A. (2021). *Doubly latent multilevel procedures for organizational assessment and prediction. Journal of Business and Psychology*.
https://doi.org/10.1007/s10869-021-09736-5

Abstract

Organizational research has a rich tradition of multilevel research anchored in a long-standing recognition that conclusions obtained at the individual level of analysis cannot be expected to generalize at other levels. Doubly latent multilevel methods, which provide the way to combine multilevel analyses and structural equation models into a single analytic framework, enrich this tradition. Yet, some critical technical considerations have yet to be systematically integrated in applied research. This article aims to introduce organizational researchers to the estimation of doubly latent multilevel models while taking into account: (a) the need to systematically assess the multilevel measurement structure of the constructs included in one study; (b) the various types of measurement errors that can be controlled for (rather than simply estimated) as part of these models; (c) the importance of relying on a clear understanding of contextual versus climate constructs given their role in centering decisions. These issues are illustrated by an investigation of the multilevel relations between psychological empowerment, psychological health, and turnover intentions among a large sample (N = 5875 employees nested within 49 work units) of Canadian Defence employees.

Keywords. Multilevel; Doubly Latent; Latent Measurement; Latent Aggregation; Multilevel Measurement; Context; Climate; Centering; Multilevel Mediation; Psychological Empowerment; Psychological Health; Turnover Intentions.

Conclusions obtained at one level of analysis (e.g., the individual) cannot be expected to generalize at another level of analysis (e.g., the work unit). This recognition is deeply ingrained in the organizational sciences (e.g., Gonzàlez-Romà & Hernàndez, 2017) where it has generated a rich multilevel research tradition (e.g., Bliese et al., 2019; Gully, Incalcaterra, Joshi, & Beaubien, 2002; Mathieu & Chen, 2011; Maynard, Gilson, & Mathieu, 2012). This tradition has long been anchored in a strong theoretical understanding of the ways distinct constructs need to be conceptualized and operationalized in multilevel research (e.g., Chan, 1998; Klein & Kozlowski, 2000).

Doubly latent multilevel confirmatory factor analyses (DL-MLCFA) and doubly latent multilevel structural equation modeling (DL-MLSEM) help to enrich this rich multilevel research tradition. These types of analytic models are starting to be more frequently implemented in organizational research following introductions by Preacher and colleagues (Preacher, Zhang, & Zyphur, 2011, 2016; Preacher, Zyphur, & Zhang, 2010; Zhang, Zyphur, & Preacher, 2009) to their use for multilevel tests of mediation or moderation. Yet, through their focus on mediation/moderation, these introductions did not fully cover some more technical elements associated with the estimation of these models, which have rather been introduced in the educational psychology (Marsh et al., 2012; Morin, Marsh, Nagengast, & Scalas, 2014) or statistical (Lütdke et al., 2008, 2011; Marsh et al., 2009) literature.

DL-MLCFA and DL-MLSEM are powerful statistical tools that respectively help inform measurement and prediction. Their advantages for organizational research are numerous and should be relatively well understood in our research tradition, and yet they remain generally underused. More precisely, doubly latent procedures are doubly important for organizational research as they combine two equally important analytical traditions: (a) Confirmatory Factor Analyses (CFA) and Structural Equation Modeling (SEM), and (b) Multilevel Analyses (MLA). Thus, on the one hand, they provide a way to assess relations among latent constructs corrected for unreliability (as in CFA and SEM), making it possible to estimate critically important theoretical mechanisms with a greater level of precision. On the other hand, they provide a way to disaggregate and contrast effects occurring across levels of analyses. As such, their combination makes it possible to assess any relation occurring across levels of analyses (i.e., individual, group and organization, or occasion, individual and group) in a way that is corrected for various forms of unreliability, making them useful for any multilevel investigation where constructs are measured with some degree of unreliability. Given the widespread reliance on questionnaire data in organizational research, it is hard to downplay this advantage. Furthermore, these models also provide a way to assess group-level constructs in a way that incorporate one additional layer of correction for unreliability related to the inter-rater agreement between the various individual ratings used to assess the group reality. This second layer of correction is why these models are called doubly latent: They provide a way to assess latent constructs using multiple items while correcting for inter-item reliability (latent measurement), as well as to assess group-level constructs using multiple individual raters while correcting for inter-rater agreement (latent aggregation).

This article seeks to introduce organizational researchers to the implementation of DL-MLCFA and DL-MLSEM and offer precisions and guidelines regarding technical aspects related to doubly latent estimation that have yet to be routinely integrated and understood in the organizational sciences. More precisely, we first provide a quick overview of the advantages of CFA/SEM, MLA, and their combination. Second, we briefly summarize the nature of the data set used to illustrate doubly latent procedures and their advantages. Third, we address key considerations that uniquely apply to doubly latent estimation procedures, rather than to the separate estimation of CFA/SEM and MLA models. These considerations, related to doubly latent multilevel measurement, measurement errors, and centering are presented and followed by an illustration. Fourth, we summarize the key benefits of doubly latent estimation procedures and illustrate how alternative models relying on partial corrections (e.g., latent measurement or latent aggregation) can be estimated. Fifth, we discuss the methodological and practical implications of doubly latent models and, in doing so, provide a summary set of guidelines for their estimation.

Doubly Latent Multilevel Models

Confirmatory Factor Analyses (CFA) and Structural Equation Modeling (SEM). Most organizational research questions involve increasingly complex theoretical models involving chains of relations among variables, multi-group comparisons, or longitudinal analyses, which can all be estimated within a single step (thus reducing risks of making type 1 errors) in the SEM framework (e.g., Bollen, 1989). More importantly, although organizational research sometimes involves

objectively measured data, the bulk of our research relies on questionnaire data where latent (unobservable) constructs (commitment, empowerment, leadership, etc.) are measured via a series of imperfect indicators (i.e., questionnaire items). Indeed, scores obtained on each of these indicators include some degree of random measurement error. For this reason, their combination into a single observed score on the construct of interest also incorporates some of this unreliable variance, leading to an underestimation of predictive relations (e.g., Nunnally & Bernstein, 1994). This issue is even more pronounced in research involving moderators (i.e., tests of interactions where the effect of a predictor on an outcome is assumed to vary as a function of a moderator; Marsh et al., 2013) or congruence between two sets of ratings (such as in tests of person-organization value fit or work-life discrepancies; Cheung, 2009). In this regard, the key advantage of SEM is to make it possible to estimate predictive models from latent variables (or factors), directly estimated from their indicators and entirely corrected for unreliability. Indeed, in any type of latent factor model, unreliability is explicitly absorbed into the uniquenesses (i.e., residual variance) of the indicators, and thus extracted from the latent construct. A core implication of this ability to control from measurement error is that predictive SEM models should be built upon adequately specified measurement models (CFA).

Multilevel Analyses (MLA). Organizational research is naturally multilevel in nature, and thus greatly benefits from MLA. Indeed, whereas our research methodology often involves measures taken from a sample of employees, these employees are also members of workgroups, nested within different organizations. Beyond the well-established, and purely statistical, need to account for this to obtain unbiased estimates of standard errors (e.g., Hox, 2010), this multilevel structure (individuals, workgroups, organizations) is also very often intimately related to our research questions so that ignoring it is likely to lead to completely erroneous conclusions (Gonzàlez-Romà & Hernàndez, 2017). For instance, leadership research naturally involves effects occurring at the workgroup level, where one supervisor's behaviors are assumed to equally influence all employees placed under his or her supervision (Chun et al., 2009; Hoffman, Bynum, Piccolo, & Sutton, 2011). This inherently group-level question can easily be accompanied by the simultaneous consideration of an individual-level research question focusing on the complementary influence of the unique individual interactions taking place between each employee and the supervisor. Similar considerations apply to research focusing on work climate (e.g., Chen, Liu & Portnoy, 2012; Morrison, Wheeler-Smith, & Kamdar, 2011), organizational culture (e.g., Schneider, Gonzàlez-Romà, Ostroff, & West, 2017; Huhtala, Tolvanen, Mauno, & Feldt, 2015), teamwork (e.g., Chen et al., 2013; Hu & Liden, 2011; Mathieu et al., 2017), or even change readiness (e.g., Rafferty, Jimmieson, & Armenakis, 2013). Across all of these domains, a focus on inter-individual differences in experiences and perceptions can easily be combined with a focus on macro-level issues that would be completely missed using a single level of analysis. For instance, in research on organizational citizenship behaviors, Schnake and Dumler (2010) reinforced that, although these individual behaviors can be impacted by a variety of individual (e.g., commitment), group (e.g., leadership), and organizational (e.g., work context) characteristics, it is mainly at the group and organization level that the beneficial outcomes of these behaviors can be observed. Importantly, even for research focusing on individual constructs (e.g., commitment, motivation, or passion for work: Fedor, Caldwell, & Herold, 2006; Gagné et al., 2019; Liu, Chen, & Yao, 2011), acknowledging that workgroup or organizational processes can influence these constructs requires the ability to properly disaggregate the effects occurring at these levels (Zhang et al., 2009).

Doubly Latent Multilevel Models (DL-MLCFA and DL-MLSEM). Although it is always useful to refresh our collective memories regarding the utility of CFA, SEM and MLA, they are nowadays routinely applied in organizational research. Unfortunately, despite their widespread application, the bulk of research published to date still relies on CFA/SEM *or* MLA, rather than on CFA/SEM *and* MLA. As noted above, this lack of integration of these two equally rich analytic traditions seems to stem from a lack of familiarity of organizational researchers with the more recently developed dual DL-MLSEM framework, which caries its own load of complexity. First, extending the estimation of latent variables across two levels of analyses poses some challenging questions related to whether, and how, the well-established measurement structure typically identified using individual ratings will generalize to the group-level. Second, combining CFA/SEM with MLA offers researchers the possibility to measure, and to control for, three distinct types of measurement errors whose combined effects, rather than simply resulting in the underestimation of associations among constructs, is also likely to generate artificial relations (Marsh, Seaton et al., 2010). Finally, although DL-MLCFA and

DL-MLSEM analyses are able to disaggregate ratings provided by employees into their individual and group components, at least two distinct types of ratings can be provided by employees. On the one hand, employees can be asked to report on their own reality (e.g., motivation, relationships with their supervisors). On the other hand, they can be asked to directly rate their workgroup reality (e.g., leadership behaviors, group climate). These apparently simple measurement decisions, however, have important implications for the interpretation of DL-MLSEM results. The goal of this article is to introduce readers to these considerations.

We note, however, that doubly latent models are complex, and require a good preliminary understanding of the principles underlying the separate estimation of MLA and CFA/SEM analyses. In the present article, we assume that readers are already familiar with these two types of models, as well as with their estimation in the Mplus (Muthén & Muthén, 2019) statistical package that we use for illustration purposes. Readers interested in learning more about these foundational analytic models are referred to: (a) Hox (2010) for a generic introduction to MLA; (b) Heck and Thomas (2015) for an introduction to the implementation of MLA using Mplus; (c) Kline (2016) for a generic introduction to CFA/SEM; (d) Geiser (2012) for an introduction to the implementation of CFA/SEM using Mplus. In addition, Morin et al., (2014) provide a user-friendly introduction to DL-MLCFA and DL-MLSEM which can provide a strong foundation upon which to anchor the concepts covered in the present article. Finally, Bliese Maltarich and Hendricks (2018) and Gonzàlez-Romà and Hernàndez (2017) both provide good introductions to multilevel analyses specifically focused on organization research.

Data Illustration: Psychological Empowerment, Psychological Health, and Turnover Intentions

To help readers to maximally connect with this introduction, we rely on a worked example focusing on the multilevel measurement of, and assessment of relations between, psychological empowerment (PE), psychological health (PH), and turnover intentions (TI). These constructs were measured as part of organizational assessment procedures conducted within the Canadian Armed Forces/Department of National Defence (CAF/DND). We rely on archival data from a convenience sample of 5,875 unique CAF/DND employees (64% males, 17% females, and 19% did not indicate their sex; 70% military, 14% civilians, and 16% did not indicate their status), working within 49 work units (including 15 to 387 employees, M = 119.898; SD = 83.405). The average response rate of employees within work units was 68%. These participants completed the Unit Morale Profile version 2.0 (UMP2; Ivey, Michaud, Blanc, & Dobreva-Martinova, 2018) between April 2014 and May 2017. Most participants (76%) completed the English version of the UMP2, whereas the others completed the French version. The UMP2 was approved by the CAF/DND Social Science Research Review Board. All respondents provided informed consent and were ensured of the confidentiality of their responses. As part of the UMP2, TI were measured with a single subscale, whereas four subscales were used to assess the core components (meaning, impact, self-determination, and competence) of PE (Spreitzer, 1995). Finally, because the main focus of the UMP2 was on PH, this construct was measured via subscales focusing on employees' burnout (disengagement and emotional exhaustion), job engagement (cognitive, physical, and emotional engagement), morale, and psychological distress.

Although our objective remains one of methodological illustration, readers showing a theoretical interest in this analytic example will find a more extensive presentation of the measures used in this study (Section 1), a theoretically-driven introduction to our analyses (Section 2), and a theoretically-driven discussion of our results (Section 3) in the online supplements. Furthermore, in alignment with our methodological objective, the literature review presented in Section 2 explicitly highlights whether and how previous research has considered, or failed to consider, the three methodological issues that form the core of the present article.

All analyses presented in this article were conducted using the Mplus 8.3 statistical package (Muthén & Muthén, 2019), using the Maximum Likelihood-Robust estimator (MLR, showing robustness to non-normality and nesting within work units) and Full information Maximum Likelihood (FIML) estimation procedures to handle missing data (Enders, 2010). Although most models were explicitly estimated as multilevel models (L1: individuals; L2: work units), preliminary measurement models were also estimated as single level models (at the individual level) while controlling for participants' nesting into work units with the Mplus design-based correction procedures (Asparouhov, 2005). Across models, goodness-of-fit was assessed with the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the Root Mean Square Error of Approximation (RMSEA), and an inspection of parameter estimates. We also report the robust χ^2 test statistic. Typical interpretation

guidelines (Hu & Bentler, 1999; Marsh, Hau, & Grayson, 2005) suggest that excellent and adequate model fit are reflected by CFI and TLI values greater than .95 and .90, and RMSEA values lower than .06 and .08. For all analyses conducted in the present article, we provide (as part of online supplementary materials) instructions and annotated Mplus (Muthén & Muthén, 2019) files for the estimation of similar analyses. At the time of writing this article, Mplus remained the most flexible (and user-friendly) framework for the estimation of this type of model, although recent developments in the *R* statistical platform are also promising (Rockwood, 2020).

Multilevel Assessment and Prediction: Challenges and Solutions

Multilevel Assessment of Conceptual Distinctiveness Using DL-MLCFA

Conceptual Distinctiveness of Multilevel Constructs

Whenever subscales taken from distinct instruments are combined into a single study, it is a commonly agreed upon best practice to demonstrate that these subscales can be differentiated in the sample under study. This verification is typically conducted with preliminary factor analyses such as CFA. This verification is important to ensure that the measures work well psychometrically and that there is no multicollinearity among them. This verification is also conceptually important because even constructs that are well-differentiated conceptually can be hard to differentiate empirically (Casper, Vaziri, Wayne, DeHauw, & Greenhaus, 2018; Marsh, Craven, Hinkley, & Debus, 2003; Weidman, Steckler, & Tracy, 2017). Although this could happen because of a lack of conceptual clarity, it may also simply reflect the fact that well-functioning individuals tend to function well across a variety of domains. Importantly, group-level (L2) aggregations tend to be characterized by even higher correlations than individual ratings. Thus, even when individual group members are able to differentially perceive distinct facets of the group functioning, their aggregated perceptions may reflect to a far greater extent the global functioning of the group across dimensions rather than the distinctive nature of these dimensions. For instance, Morin et al. (2014) found that three facets classroom climate (mastery, challenge, and caring), which were well differentiated at the level of individual students' perceptions, were impossible to distinguish at the classroom level. Just like well-functioning individuals, well-functioning groups tend to function well across indicators. Importantly, better controls for measurement errors (which we address in the next section) also tend to generate higher correlations, making it even more important to ascertain conceptual distinctiveness when studying multilevel phenomena. Unfortunately, very few organizational measures have been systematically validated for use in multilevel contexts (e.g., team efficacy: Arthur, Bell, & Edwards, 2007; organizational processes and practices: Shortell et al., 1991).

Higher-Order Models and Bifactor Models

Modern factor analytic techniques make it possible to "have your cake and eat it to" when conceptually differentiated constructs are found to overlap empirically (e.g., Morin, Myers, & Lee, 2020; Morin et al., 2014), and can easily be extended to multilevel analyses. Indeed, higher-order factor models and bifactor models provide a way to disaggregate ratings obtained across multiple dimensions into a global component, reflecting the variance shared across all dimensions, and specific components reflecting what is uniquely associated with each subscale over and above the estimation of this global component. In a higher-order model, first-order factors are estimated directly from item ratings, the variance shared among these factors is used to estimate a second-order factor, and the residuals associated with the first-order factors reflect their unique variance not shared with the other dimensions. In contrast, bifactor models directly estimate the global and specific factors from the items themselves, leading to a natural disaggregation of the variance shared across all items as part of a global factor (G-factor), and of the variance shared among items forming a subscale but not explained by the global factor as part of specific factors (S-factors). Bifactor models, albeit less parsimonious, are more flexible in allowing for direct estimation of associations between items and latent constructs (Gignac, 2016; Reise, 2012). Both approaches can easily be implemented at L1 and L2 with DL-MLCFA and DL-MLSEM (e.g., Morin et al., 2014).

Importantly, higher-order factor models are limited by the inclusion of a stringent proportionality constraint in which the ratio of item-level variance explained by the higher-order factor relative to that explained by the first-order factor is forced to be constant across all items linked to the same first-order factor (i.e., all items taken from the same subscale; Gignac, 2016; Morin et al., 2020). Although this constraint only has a limited impact on the way the results from both types of models are interpreted, it does has a strong statistical impact on model estimation, making higher-order models far

less likely to achieve a satisfactory representation of the data. In addition, whereas the global factor estimated as part of higher-order or bifactor models can be interpreted as reflecting the variance shared among all indicators[1], the meaning of the specific factors differs across models. In a bifactor model, the S-factors reflect the variance shared among the items that is not already explained by the G-factor, thus representing the extent to which scores on these dimensions deviate from scores on the G-factor. In contrast, in a higher-order model, the first-order factors can be interpreted as any other factor (i.e., reflecting the variance shared among its indicators). Indeed, these first-order factors include both the variance explained by the G-factor and the variance unique to the first-order factor, thus creating interpretational difficulties and redundancies in models where both the higher-order and the first-order factors are used simultaneously (Morin, Boudrias, et al., 2017). In a higher-order factor model, a direct estimate of the variance uniquely associated with the first-order factor (thus corresponding to the meaning of a bifactor S-factor) is found in the disturbance (i.e., residual) of the first-order factor.

In practice, our recommendation is to always start with the estimation of a typical correlated factor CFA model. Then, alternative bifactor solutions should be considered in two different situations and contrasted with the initial correlated factor solution. First, bifactor alternatives should be investigated whenever there is a theoretical or empirical (i.e. based on the results from previous research) rationale suggesting that such a model might be appropriate, as it is the case, for example, in research focusing on psychological health and well-being (Morin, Boudrias, et al., 2016, 2017), motivation (Howard et al., 2018), or burnout (Mészáros et al., 2014). Second, bifactor alternatives should also be investigated whenever factor correlations estimated at L1 or L2 from the initial correlated factors CFA model are high enough to suggest conceptual redundancies or to generate multicollinearity, as long as it makes theoretical, empirical, or even logical sense to combine these various measures to assess one global (G-) factor (i.e., if it is possible to label this G-factor in a way that makes sense). In situations where this approach makes no sense, then the researcher will be left with the choice between eliminating some of the overlapping variables, combining them into a single construct, or relying on partial bifactor solution in which only a subset of these variables are used to estimate the G-factor. Due to the aforementioned proportionality constraint and interpretational challenges, we only recommend the use of higher-order models when they are clearly supported by theory or prior research, such as it is the case for PE (Seibert et al., 2011). Even then, higher-order factor models should routinely be contrasted with bifactor models before making a final decision.

Multilevel Measurement Isomorphism

A final measurement consideration is related to the measurement *isomorphism* of L1 and L2 constructs estimated from the same ratings. Measurement *isomorphism* occurs when constructs have the same measurement structure, with the same factor loadings, at both levels (e.g., Bliese, Chan, & Ployhart, 2007; Metha & Neale, 2005; Tay, Woo, & Vermunt, 2014). From an interpretational standpoint, *isomorphism* means that the matching L1 and L2 factors can be interpreted in a similar manner. One important caveat is that *isomorphism* considerations differ slightly in doubly latent procedures (i.e., when the same ratings are used simultaneously to assess the individual and the group reality using latent aggregation procedures) relative to other types of multilevel models when the latent constructs used to reflect the individual and group reality are obtained from different sets of items.

First, when constructs are uniquely and directly assessed at the group (e.g., using supervisors self-reports of their leadership styles) or individual (e.g., employees' ratings of their own individual self-efficacy) level and the analyst has no interest in aggregating individual ratings at the group level (e.g., team efficacy) or in using matching individual measures of the construct assessed at the group level (e.g., employees' perceptions of their supervisor's leadership style), then *isomorphism* is not relevant.

Second, when distinct sets of measures are used to assess conceptually-related constructs across levels (e.g., asking supervisors to report on their leadership and employees to report on their supervisor's leadership, or asking the employees to separately assess their individual and group efficacy) and when these measures present a similar structure (using matching dimensions and items), then *isomorphism* might become a consideration. In this situation, the core consideration is theoretical, and related to whether one has empirical or theoretical reasons to expect *isomorphism* (e.g., Chan,

[1] This is exactly the case in a bifactor model where all items are directly used to define the global factor, and approximately the case in a higher-order model where the global factor reflects the variance shared among first-order factors, themselves reflecting the variance shared among subsets of indicators.

1998; Chen, Bliese, & Mathieu, 2005; Tay et al., 2014). In these situations, beyond these theoretical considerations, the ability to achieve a measurement that is equivalent (with a matching factor structure and equal factor loading) will also yield some statistical benefits in terms of parsimony.

The third situation is specifically related to the estimation of doubly latent constructs, where latent aggregation procedures are used to obtain an estimate of separate constructs reflecting the group and individual reality obtained from the same set of items completed by the employees. In this situation, isomorphism has clear advantages for most applications. Essentially, *isomorphism* allows one to directly conceptualize the L1 latent construct as being a random variable with L2 variability (Metha & Neale, 2005). In other words, by equating the metrics of measurement (i.e., demonstrating an equivalent measurement models with equal factor loadings across levels), it becomes possible to directly compare the construct and the way it relates to other variables across levels. Later, we differentiate two types of doubly latent constructs. Contextual variables are group level variables formed by the aggregation of meaningful individual ratings (e.g., such as when employees' gender is measured at L1 and used to obtain an estimate of the gender composition of the work group at L2). In this situation, *isomorphism* is necessary to obtain proper estimates of contextual effects due to the need to subtract L1 effects from their L2 counterpart to achieve a proper estimate of the contextual effect (these considerations are detailed later in our section on centering). In contrast, climate variables are obtained when employees are directly asked to rate the L2 group reality. In this situation, the L1 component of these ratings simply reflects inter-individual deviations in ratings of this L2 reality and might be, or not, of interest to researchers. Whenever researchers are interested in interpreting L1 components as reflecting inter-individual variations in ratings of this L2 reality, then *isomorphism* is also necessary. Otherwise, the metric of these individual components will no longer match that of the group reality, precluding any meaningful interpretation of the L1 component of climate ratings. In contrast, when researchers are not interested in the interpretation of these L1 components, then *isomorphism* is not required. Finally, and perhaps more importantly, Lütdke et al. (2008, 2011) have shown that imposing *isomorphism*, even when it is not fully supported by the data, helps to stabilize the model estimation process and to obtain more accurate parameter estimates. For this reason, we recommend testing for *isomorphism*, and even imposing it unless the results clearly fail to support it (i.e., drastic decrease in model fit), for the assessment of any type of doubly latent construct.

Data Illustration – Step 1: Measurement Model Estimation

In any DL-MLSEM application, the first analytical step is to verify the multilevel measurement structure and measurement *isomorphism* (i.e., metric invariance, or equality of the factor loadings across levels) of the various latent constructs used in any specific study via a combination of single level CFA and multilevel DL-MLCFA. Importantly, the results from these preliminary measurement models will provide the information required to calculate the various types of reliabilities that will be covered in the next section. In the present example, previous research does provide a strong level of empirical support for the superiority of a higher-order representation of PE (e.g., Seibert et al., 2011) and for a bifactor representation of PH (Laguna et al., 2019; Morin, Boudrias et al., 2016). We thus rely on DL-MLCFA and DL-MLSEM to verify whether this structure will generalize to the UMP2 assessment proposed by CAF/DND and across levels of analyses. Given the complexity of the models considered, these preliminary verifications were conducted on a single type of construct at a time (TI, PE, and PH). The goodness of fit of all models estimated in this study is reported in Table 1.

Turnover Intentions. The simplest model was for TI, involving a single factor model at both levels. Because a single level CFA model including only three indicators is just identified and thus always has perfect fit, goodness of fit information was not reported in Table 1 for the single level TI model as well as for the multilevel TI model with factor loadings freely estimated across levels. However, the parameter estimates from these models revealed strong and satisfactory factor loadings, suggesting no specific issue with these models. Likewise, the *isomorphic* multilevel measurement model for TI was able to achieve a satisfactory level of fit to the data, resulting in strong and equivalent factor loadings across models and levels (.661 to .927, M_{λ} = .774; see the top section of Table S1 in Section 4 of the online supplements). This evidence of *isomorphism* supports the idea that the measurement of TI follows a similar structure across levels. Annotated syntax files used to estimate this model at L1 only, then at L1 and L2 (multilevel), and to test for measurement *isomorphism* across levels, are presented in Section 5 of the online supplements (this section also includes some generic notes relevant to all syntax files provided in the supplements).

Psychological Empowerment. For the PE construct, a CFA model in which four correlated factors reflecting the four components of PE (self-determination, meaning, and impact, and competence) was first estimated at both levels, and contrasted with the a priori higher-order factor model in which these four dimensions were specified as first-order indicators of a higher-order PE factor (Seibert et al., 2011). A bifactor representation of PE was also estimated in which each item was simultaneously used to define one global PE factor in addition to the a priori self-determination, meaning, and impact, and competence factors, and all factors were set to be orthogonal according to typical bifactor representations (Morin et al., 2020). We reinforce here that this orthogonality is not different from the orthogonality of the first-order factors estimated in a higher-order model. In these models, three a priori correlated uniquenesses were added at the individual level (L1) to control for the negative wording of three of the self-determination items (Marsh, Scalas, & Nagengast, 2010). Annotated syntax files used to estimate these alternative models at L1 only, and the higher-order models at L1 and L2 (multilevel) are presented in Section 6 of the online supplements.

As shown in Table 1, all three models (correlated factors CFA, higher-order CFA, and bifactor-CFA) were able to achieve a similar and satisfactory level of fit to the data in the single level models. However, the superiority of the correlated factors CFA and higher-order CFA, relative to that of the bifactor-CFA, was apparent in the DL-MLCFA models. Both the correlated factors CFA and the higher-CFA were able to achieve comparable and satisfactory fit to the data, and evidence of *isomorphism* (equality of the factor loadings) across levels. The most parsimonious higher-order factor model was retained, supporting our a priori expectations. The parameter estimates from this model are reported in Table S1 of the online supplements and, lending further support to this higher-order solution, revealed a well-defined higher-order factor (with loadings ranging from .349 to .917 across levels and solutions, M_{λ} = .691), accompanied by equally well-defined competence (.782 to .989, M_{λ} = .851), meaning (.931 to .999, M_{λ} = .963), self-determination (.355 to .950, M_{λ} = .687), and impact (.775 to .983, M_{λ} = .900) first-order factors. Essentially, this solution indicates that PE encompasses four reasonably differentiated components (competence, meaning, self-determination, and impact), which themselves form a single overarching factor. Furthermore, this solution also suggests that this measurement structure generalizes well when individual ratings are aggregated to the group-level.

Psychological Health. For PH a CFA model in which seven correlated factors representing participants' levels on each PH component (disengagement, emotional exhaustion, cognitive engagement, emotional engagement, physical engagement, morale, and psychological distress) was first estimated at both levels. This model was then contrasted to our a priori (Morin, Boudrias et al., 2016) bifactor solution where each item was allowed to load on both one global PH factor and to their a priori factor, with all factors set to be orthogonal[2]. Given this a priori support for a bifactor representation, we had no reasons to estimate higher-order representations of PH (subsequent verifications confirmed the inadequacy of higher-order models for PH). All of these models included two a priori correlated uniquenesses at the individual level (L1) to control for the parallel wording of two pairs of items from the psychological distress scale (Marsh et al., 2013). Annotated syntax files used to estimate these alternative models at L1 only, and the bifactor models at L1 and L2 (multilevel) are presented in Section 7 of the online supplements.

As shown in Table 1, both the correlated factors CFA model and the bifactor model were able to achieve a comparable and satisfactory level of fit to the data in single level models. The pattern of results was slightly more complex with the DL-MLCFA results. These results first showed that, although the correlated factor CFA model was unable to achieve a proper level of fit initially, the imposition of equal factor loadings across levels (reported by Lütdke et al. [2008, 2011] as having a stabilizing impact on model estimation) made it possible to achieve acceptable fit for this solution. Yet, this solution revealed factor correlations reaching .873 at L1 and .953 at L2, suggesting a substantial level of conceptual overlap across factors. The bifactor model achieved a satisfactory level

[2] Following Morin, Boudrias et al. (2016) we also estimated alternative bifactor solutions including two correlated global factors representing participants' global levels of psychological well-being (cognitive engagement, emotional engagement, physical engagement, morale) and distress (disengagement, emotional exhaustion, psychological distress). However, these alternative models resulted in a similar level of fit to the data than their counterparts including a single global factor, and in the estimation of two G-factors that were so highly correlated at L1 (r = -.838) and L2 (G-factor r = -.932) to suggest conceptual overlap.

of fit to the data, higher than that of the correlated factor model when factor loadings were free across levels, and comparable to it when *isomorphism* was imposed. This model was thus retained, supporting the idea that the measurement structure of PH can be generalized at the group level.

The parameter estimates from this model are reported in Table S2 of the online supplements. When interpreting bifactor results, it is important to keep in mind that in these models, the true score (i.e., reliable) variance present in each item is divided across two factors (G-factor and S-factor) so that factor loadings (and reliability estimates) are typically smaller (Morin et al., 2020). Likewise, it is typical for bifactor model solutions to reveal some items that present a dominant association with one of these two factors, which should not be taken as evidence of any sort of problem with the item, simply as evidence that this item is a better indicator of one of these two layers of measurement (Morin et al., 2020). Similarly, although some S-factors should retain a significant level of specificity to support of bifactor operationalization, it is not necessary for all S-factors to retain specificity and bifactor parameters estimates are notably robust to vanishing S-factors (Morin et al., 2020). With this in mind, these results revealed a well-defined G-factor characterized by positive factor loadings from all of the well-being indicators across models and levels (.173 to .951, M_{λ} = .624) and negative factor-loadings from the distress indicators (-.471 to -.956, M_{λ} = -.699). This G-factor can thus be taken to reflect participants' global levels of PH across all indicators. Similarly, the well-being S-factors all appeared to retain a substantial amount of specificity once the G-factor was taken into account: Morale (.264 to .604, M_{λ} = .482), physical engagement (.493 to .828, M_{λ} = .711), emotional engagement (.215 to .573, M_{λ} = .448), and cognitive engagement (.308 to .810, M_{λ} = .624). In contrast, two out of three distress S-factors appeared to retain only a minimal amount of specificity (emotional exhaustion: .126 to .469, M_{λ} = .274; disengagement: .160 to .602, M_{λ} = .320), although this was not the case for the psychological distress S-factor (.140 to .676, M_{λ} = .443). This suggests that emotional exhaustion and disengagement show strong alignment with global levels of PH for most participants, thus retaining only a limited amount of specificity beyond this global level. In contrast, all other S-factors (moral, physical engagement, emotional engagement, cognitive engagement, and psychological distress factors) retain a meaningful level of specificity. These S-factors thus provide a direct reflection of the extent to which participants' scores on these dimensions deviate from their global levels of PH.

Complete Model. When these three measurement models (single factor for TI, higher-order for PE, and bifactor for PH) were combined into a single model, the resulting solution was also able to achieve an acceptable level of fit to the data (see Table 1) and was thus retained for predictive purposes. Multilevel latent factor correlations obtained as part of this model are reported in Table S3 of the online supplements and are consistent with between-constructs associations occurring mainly at the levels of the global/higher-order relative to the specific/first-order factors. The syntax used to estimate this model is entirely reproduced in Section 8 of the online supplements.

Multilevel Sources of Measurement Error

When ratings taken at the individual level are used to assess individual (L1) and group (L2; i.e., unit, group, etc.) constructs, three distinct forms of measurement errors need to be taken into account (Lütdke et al., 2011; Marsh et al., 2009) before proceeding to multilevel predictive analyses.

Inter-Item Reliability in Participants Ratings

The first type of measurement error refers to the reliability of the ratings provided by the participants themselves (i.e., the ratio of true score variance to total variance present at the item level: Nunnally & Bernstein, 1994). Reliability assessment considers that the variance shared among the items forming a scale reflects reliable (i.e., true score) variance, and that the variance uniquely associated with each item (not shared with the other) reflects a combination of random measurement error and item specificities. This type of measurement error is located at L1 and controlled for as part of the items' uniquenesses in confirmatory factor analyses (CFA) and structural equation models (SEM), which assess relations untainted by this first source of measurement error (e.g., Bollen, 1989).

Inter-Item Reliability in Collective Ratings

A second and similar source of measurement error is present at L2 (Geldhof, Preacher, & Zyphur, 2014; Lütdke et al., 2011). The combination of items to form a scale at the individual level (L1) is likely to reflect a specific amount of measurement error that is distinct from the measurement error present when ratings of these same items are aggregated at L2 to obtain group-level construct scores. This second source of measurement error is likely to be smaller than its L1 counterpart when estimated as part of DL-MLCFA or DL-MLSEM models in which L1 measurement error is already controlled

for. Yet, most multilevel applications rely on scale scores that are manually calculated at both levels (i.e., reflecting the mean of the items forming a scale at L1, which is then averaged at L2). In these non-latent analyses, both sources of measurement errors are likely to affect results.

Calculating Inter-Item Reliability at Both Levels

The first two sources of measurement error, related to "inter-item" agreement in the estimation of latent constructs, are assessed as part of traditional analyses of reliability. In this regard, Geldhof et al. (2014) recommended relying on the omega (ω) coefficient of composite reliability (McDonald, 1970), which can be calculated separately at the individual (ω_{L1}) and group (ω_{L2}) level as

$$\omega = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \left(\sum_{i=1}^{k} \theta_{ii}\right)}$$

where λ_i reflect the standardized factor loadings associated with the i^{th} item (out of k items) on a specific factor, and θ_{ij} are the standardized item uniquenesses, thus providing a direct estimate of the ratio of true score variance on the total variance present at the item level. Importantly, ω has also been shown to be an appropriate indicator of reliability for bifactor measurement models, with the simple interpretation caveat that bifactor ω will tend to be smaller as construct-relevant variance (expressed in the factor loadings) are divided across G- and S- factors (Morin et al., 2020).

Inter-Rater Agreement in Collective Ratings

The third source of measurement error is related to the agreement between participants in their rating of the group level (L2) reality (Bliese et al., 2019; Croon & van Veldhoven, 2007; Lütdke et al., 2008). Just like items in a scale can be considered to be alternative indicators of the same construct, each member of a collective can be considered to be alternative indicators (i.e., raters) of their collective reality. This form of error has alternatively been referred to as homogeneity/consensus (or lack thereof; e.g., Chan, 1998; Klein & Kozlowski, 2000), sampling error (e.g., Gonzàlez-Romà, & Hernàndez, 2017), and inter-rater reliability (or error; e.g., Gagné et al., 2019).

Calculating Inter-Rater Agreement

This final source of measurement error, related to the degree of "inter-rater" agreement in the assessment of the L2 construct can be estimated by considering two intra-class correlation coefficients (ICC1 and ICC2; Marsh et al., 2012). The more common ICC1 indicates the proportion of the total variance in rating occurring at L2 and reflects the average agreement among members of a single L2 unit in ratings of a specific construct (x), whereas the ICC2 directly reflects the reliability of the group (L2) aggregate (and can be interpreted as any other reliability estimates):

$$ICC1 = \frac{\tau_x^2}{\tau_x^2 + \sigma_x^2} \qquad ICC2 = \frac{\tau_x^2}{\tau_x^2 + \left(\sigma_x^2 / n_j\right)}$$

where τ_x^2 refers to the L2 variance, σ_x^2 to the L1 variance, and n_j to the average number of participants in each of the L2 units. Typical interpretation guidelines suggest that ICC1 values should be ideally greater than .100 but minimally greater than .050 to justify relying on multilevel procedures (e.g., Gonzàlez-Romà, & Hernàndez, 2017; Lütdke et al., 2008, 2011), whereas ICC2, ω_{L1} and ω_{L2} values should ideally be greater than .700 (e.g., Klein & Kozlowski, 2000; Morin et al., 2014).

Doubly Latent Models and Reliability

Just like CFA and SEM models estimate relations between latent constructs untainted by measurement error, DL-MLCFA and DL-MLSEM estimate relations among constructs untainted by all three types of measurement errors. Thus, DL-MLCFA and DL-MLSEM are called *doubly latent* because they provide a way to control for both sources of item-level unreliability by using *latent variables* estimated directly at the item level at L1 and L2, in combination with a *latent aggregation* process whereby agreement among individual (L1) ratings are used to control for the third source of measurement error (Marsh et al., 2009, 2012). This ability to control for three types of measurement errors is particularly important. Indeed, whereas failure to control for the first source of measurement error simply results in downwardly biased estimates of relations among constructs, failure to control for these three sources of measurement errors in multilevel analyses can lead to the estimation of non-existing relations among constructs (referred to as phantom effects: Marsh et al., 2010).

It is important to keep in mind that the rough interpretation guidelines typically used to make

sense of these coefficients, such as those provided for the ICC1 and ICC2, were initially proposed to justify the aggregation of items into scale scores, and the aggregation of participants into L2 composites (Klein & Kozlowski, 2000). These guidelines cannot be directly applied to doubly latent models, which are robust to these three types of errors and thus provide unbiased estimates even when reliability is low (Marsh et al., 2012; Morin et al., 2014). More precisely, lower than optimal reliability estimates are not as problematic in DL-MLSEM given the ability of these models to control for a lack of inter-item, or inter-rater, reliability at both levels. In fact, low reliability even reinforces the need to adopt a DL-MLSEM procedure, far more than it argues against it. This does not mean that reliability is not important, or that measures associated with very low (e.g., .200, .300) reliability can still be considered for analyses. It simply means that flexibility is called for. Very low reliability indicates that the items (ω), or group members (ICC2), have nothing in common, calling into question the appropriateness of the measures or aggregation procedures. However, whereas most applications of multilevel analyses in the organizational sciences report some form of reliability estimate for L1 constructs, and some sort of inter-rater reliability estimates for L2 aggregates, reliability estimates for L2 constructs are very rarely reported. Furthermore, these three forms of measurement error can (and should) be estimated and controlled for with modern statistical procedures such as DL-MLSEM, rather than simply discussed as part of preliminary analyses.

Data Illustration – Step 2: Reliability Estimates

In traditional (non-latent) applications of MLA (e.g., Hox, 2010), the reliability of the variables under study and the extent to which these variables present a meaningful level of variability at L2 are estimated prior to any multilevel analyses. As noted above, these considerations are not as critical in DL-MLSEM, as these models provide a way to account and control for unreliability. Importantly, in DL-MLSEM, reliability considerations can only be addressed using parameter estimates obtained from the initial DL-MLCFA model (ideally from the isomorphic model, when supported by the data). Reliability estimation thus represents the second analytical step in the estimation of DL-MLSEM. Reliability estimates were calculated using the parameter estimates from the complete *isomorphic* measurement model reported at the end of the previous section. More precisely, standardized factor loadings and uniquenesses were used to obtain estimates of composite reliability (ω, ω_{L1}, ω_{L2}), and unstandardized variance estimates were used to obtain indices of inter-rater agreement (ICC1, ICC2). Calculations used to obtain these coefficients are exemplified in Section 9 of the online supplements.

Reliability information calculated on the basis of the final retained measurement model are reported in Table 2. When we first consider inter-item reliability, these results revealed highly reliable factors across levels and models (ω = .769 to .977, *M* = .880; ω_{L1}= .765 to .976, *M* = .877; ω_{L2} = .807 to .998, *M* = .920). The only exception to this generic conclusion is related to the two distress S-factors, which were found to retain a lower level of specificity as part of the measurement models (emotional exhaustion: ω = .514; ω_{L1}= .517; ω_{L2} = .474; disengagement: ω = .437; ω_{L1}= .419; ω_{L2} = .788). This observation suggests that the results associated with these S-factors, even though obtained as part of models incorporating a correction for measurement error, should be interpreted cautiously.

In terms of inter-rater agreement, the ICC1 first revealed that all of the PE factors (.037 for competence to .077 for self-determination, and .064 for the higher-order factor), as well as the PH G-factor (.065), all presented a sufficient level of between-group variance to support multilevel analyses (especially in the context of the large sample size available for this study at L1 and L2). In contrast, the PH S-factors, as well as TI, appeared to retain a slightly lower level of between-group variance (ranging from .006 for the cognitive engagement S-factor to .077 for the disengagement S-factor and .024 for TI). Yet, and providing further support for the reliance on multilevel analyses, the inter-rater reliability of the L2 constructs was satisfactory for most constructs (ICC2 = .600 to .909, *M* = .798), with the sole exception of the cognitive engagement S-factor (.420). Taken together, these results suggest caution in the interpretation of L2 results associated with the cognitive engagement S-factor.

Multilevel Climate or Context and Centering Implications

Centering is critical to the interpretability of multilevel models. Yet, it is also a highly technical issue (e.g., Enders & Tofighi, 2007) that remains widely misunderstood in organizational research, and that poses unique additional challenges in the context of doubly latent estimation procedures. Beyond these technicalities, however, centering decisions remain intimately related to the nature of the constructs one seeks to estimate. In the upcoming section, we thus attempt to clarify centering decisions in a way that is both not technical, and clearly anchored in the type of construct

being assessed, in a way to provide clear and practical guidance to applied researchers.

Centering

Centering involves recoding scores to obtain an interpretable 0 (Enders & Tofighi, 2007; Gonzàlez-Romà, & Hernàndez, 2017). Grand-mean centering involves rescaling the variables by subtracting the sample's mean so that zero comes to reflect the sample average. Group-mean centering involves rescaling variables by subtracting the group (L2 unit) mean so that individuals' scores come to reflect within-group deviations. For constructs that are only assessed at L2 (such as using official records to assess group productivity), then there is no individual variation to consider and grand-mean centering becomes the only option. However, when constructs contain a mixture of L1 and L2 variation, and critically when ratings taken at the individual level are used to estimate constructs at both levels, then the selection of a centering method becomes critical, and should be based on the type of construct one is trying to assess.

In multilevel research involving the L2 aggregation of L1 ratings, two types of constructs are considered. These have been referred to as: (a) involving a reflective or formative aggregation based on measurement analogies (Bliese et al., 2019; Lütdke et al., 2008, 2011; Marsh et al., 2009); (b) involving consensus or compilation based on the aggregation process (Bliese et al., 2007; Quigley, Tekleab, & Tesluk, 2007; Klein & Kozlowski, 2000), or (c) climate or contextual variables based on conceptual considerations (Gagné et al., 2019; Marsh et al., 2012; Morin et al., 2014). The key distinction between these two types of constructs is related to the referent of the rating or, more precisely, to the key determinant of these ratings.

Climate Variables

Climate variables are formed by the aggregation of individual ratings where the referent, and main driver of the rating, is located at the group (L2) level. For instance, team efficacy (when assessed with items where the team is the referent) or leadership (when employees are all asked to directly rate their manager) are climate variables naturally referring to a group-level reality, and the L1 component of these ratings simply reflect inter-individual differences in perceptions of this L2 reality. This does not mean that these differences are meaningless or unlikely to play a role in prediction. Rather, this simply means that, with climate variables, the main assumption is that ratings are primarily driven by exposure to a group reality that individuals are asked to rate so that discrepancies among members of the same L2 unit reflect differences in the perception of this L2 reality.

Contextual Variables

Contextual variables are formed by the aggregation of individual ratings where the referent, and main driver of the rating, is the individual providing the rating. For instance, gender, self-efficacy, or personality are naturally individual-level variables that, when aggregated at L2, become contextual variables. When working with contextual variables, the L1 ratings have a meaning in and of themselves, irrespective of what happens at the group level, and the L2 construct is assumed to have completely distinct implications than those of its L1 counterpart. Thus, whereas being a male or a female might have implications for individual workers, being exposed to a male-dominated or female-dominated work context may come to reflect something that is entirely distinct, and is likely to have different repercussions, than being a male or a female. When working with contextual variables, the key question then becomes to assess whether the group reality created by the aggregation of individual level variables has an effect that goes beyond, or add to, the addition of the individual effects associated with the individuals forming the group. For instance, assuming that individual employees' commitment to their organization predicts higher levels of organizational citizenships behaviors (OCBs) among individual employees describes a well-established L1 effect (e.g., Meyer et al., 2002). This effect is expected to translate to a similar group-level effect. Indeed, groups including more highly committed employees each displaying higher levels of OCBs naturally leads to higher-group averaged levels of OCBs. This group-level effect, however, is meaningless, and simply reflects the combination of individual effects. The key question, when contextual effects are considered, is rather whether combined levels of commitment create a unique group context that itself generates higher, or lower, group-levels of OCBs than what would be expected from the simple combination of employees' commitment levels.

Climate Variables: Group-Mean Centering

In terms of centering, climate constructs call for group-mean centering, where L1 ratings can directly reflect deviations from the group average, and L2 aggregates can be considered to be

completely independent (i.e., uncorrelated) from their L1 counterparts (Enders & Tofighi, 2007; Marsh et al., 2012; Morin et al., 2014). In terms of interpretation, group-mean centering clarifies that the focus is, first and foremost, on the L2 effects of aggregated perceptions, which can be directly interpreted as such. In contrast, the L1 effect directly reflects the role played by inter-individual differences in perceptions of the group reality, and typically occupies a secondary position in interpretations (e.g., Marsh et al., 2012, Morin et al., 2014).

For instance, in a study where employees are asked to directly rate the behaviors enacted by their supervisors, the L2 coefficient obtained via group-mean centering provides a direct estimate of the effects played by employee's joint perceptions of their leader's behaviors. In contrast, the L1 coefficient would directly reflect the impact played by inter-individual differences in these perceptions and needs to be explicitly interpreted as such. Although these inter-individual differences could, to some extent, reflect the fact that distinct employees may share uniquely distinct interactions with their supervisors (interpersonal relationships, differential treatment, etc.), this L1 component cannot be assumed to reflect only these differential interactions. Indeed, these differences could also reflect perceptual differences, personal biases in interpersonal perceptions, differential expectations, inter-rater (un)reliability, and so on. Thus, if the goal is to focus on interpersonal interactions, then one would need to rely on a measure explicitly focusing on these interactions (i.e., which would then have to be handled as a "contextual" measure).

Contextual Variables: Grand-Mean Centering

Grand-mean centering, in which group (L2) and (L1) variance components are not explicitly disaggregated from one another (i.e., grand-mean centered ratings still contain L1 and L2 sources of variability; Enders & Tofighi, 2007), is the appropriate approach for contextual effects. With grand-mean centering, the L1 ratings retain their own "identity", allowing for the estimation of L2 effects partialled out from, or controlled for, their L1 counterparts (Enders & Tofighi, 2007; Marsh et al., 2012; Morin et al., 2014). Thus, grand-mean centering yields L2 estimates providing a direct test of whether a contextual effect is present in the data. In terms of interpretation, the key question here is whether the context created by the combination of individual characteristics brings something more than the simple accumulation of these individual characteristics. Thus, the key component here is located at L1, where coefficients obtained using grand-mean centering directly reflect the impact of these individual characteristics. In contrast, at L2, a properly calculated contextual effect (via grand-mean centering) would come to directly reflect the extent to which these aggregated individual characteristics bring something more than the simple combination of individual characteristics.

Arguably, one of the clearest examples of contextual effects is the Big-Fish-Little-Pond effect from the field of educational psychology (e.g., Marsh et al., 2014). More precisely, whereas individual levels of achievement predict stronger academic self-conceptions among students, class levels of achievement predict lower average levels of academic self-conceptions among students from the same classroom due to various social comparison mechanisms. As a contextual effect, this phenomenon thus reveals that aggregate class levels of academic self-conceptions tend to be lower among high-achieving classroom than would be expected from the simple aggregation of student-level effects.

Doubly Latent Estimation

DL-MLCFA and DL-MLSEM models rely on an implicit group-mean centering approach (i.e., group-mean centering is automatically imposed, not as a default, but as a core part of the mathematical underpinnings of doubly latent models). This phenomenon forces analysts to manually calculate contextual effects as the difference between the group-mean centered L2 effects and their L1 counterparts (Enders & Tofighi, 2007; Marsh et al., 2012; Morin et al., 2014). Otherwise, the group-mean centered L2 effect will simply reflect the combined effect of all individuals forming the group, rather than the additional effect of exposure to a specific contextual characteristic. Because of this calculation, measurement *isomorphism* is required to calculate contextual effects (Morin et al., 2014). To be as clear as possible, any multilevel model relying on a latent aggregation procedure will involve, irrespective of the analytic commands used to estimate them, group-mean centering to the extent that even requests for a grand-mean centering approach will be ignored. This is why additional code is required to obtain proper estimates of contextual effects (see Section 11 of the online supplements).

Despite the critical importance of these considerations for DL-MLSEM applications, these specific centering guidelines are still mainly ignored in organizational research. To our knowledge, Gonzàlez-Romà, and Hernàndez (2017) are the only one who alluded to this distinction, via the

recommendation that group-mean centering should be the favored approach unless one has a reason to test contextual effects. However, this recommendation was made without providing a clear distinction between contextual and climate effects, and without the recognition that DL-MLSEM models are naturally group-mean centered.

Data Illustration – Step 3: Multilevel Predictive Model

Analyses. The last analytic steps involve estimating the a priori theoretical predictive DL-MLSEM model, ideally while contrasting it (using goodness-of-fit criteria) to alternative solutions. In the present application, the individual predictive associations between PE and PH (Maynard et al., 2012, 2013; Seibert et al., 2011), PE and TI (Maynard et al., 2012; Seibert et al., 2011), and PH and TI (Mor Barak et al., 2001; Rubenstein et al., 2018) form a well-established mediational system. In the present study, we assess whether this predictive system generalizes to the work unit level. In this example, both variables used in prediction (PE and PH) can be considered to be contextual variables, and will thus require one additional analytic step in order for their L2 effects to be properly estimated. However, to best illustrate the distinction between climate and context, we will also report the untransformed L2 effects of these contextual variables (as if they were climate variables).

The retained DL-MLCFA model was converted into a predictive DL-MLSEM assuming matching associations between constructs at L1 and L2. Given our focus on associations involving the global PE and PH constructs, the a priori predictive model (M1) allowed global PE (i.e., the higher-order factor) to statistically predict global PH (i.e., the global factor from the bifactor model) and TI. We also allowed global PH to directly predict TI, thus forming a multilevel mediation model characterized by matching indirect effects of PE on TI as mediated by PH at the individual and group level (Preacher et al., 2010, 2011).

A simplified illustration of the theoretical model tested in the present study is represented in Figure 1. As shown in Figure 1, the model assumes mediation, whereby PE is assumed to predict PH ratings, as well as TI. This second prediction is assumed to occur both directly (PE–> TI) and indirectly via the mediating role of PH (PE–> PH–> TI). This predictive model is assumed to occur both at the employee level and at the work-unit level. All constructs are illustrated by ovals to reflect the fact that they are operationalized as latent variables estimated from their items. PE is illustrated as a higher-order construct estimated from first-order self-determination, impact, meaning, and competence factors, themselves estimated from their items. In contrast, PH is represented as a bifactor model including a global PH factor and a series of specific factors (illustrated with dotted lines) estimated from the same items. A complete illustration of our analytic model, following the guidelines for the illustration of doubly latent models outlined by Marsh et al. (2012) and Morin et al. (2014) appears in Figure 2. In this figure, the complete measurement structure is illustrated separately at both levels, with the items represented in the middle to illustrate that the same items, rated by employees, are used to assess constructs at the individual and work group levels (via latent aggregation).

In a second step, we estimated a series of alternative models seeking to verify the possible additional effects of the PE first-order dimensions and of the specific PH factors (based on recommendations from Maynard et al., 2012) in the following sequence: (a) the first-order PE dimensions were allowed to predict TI (M2); (b) the specific PH factors were allowed to predict TI (M3); (c) the first-order PE dimensions were allowed to predict the global PH factor (M4); (d) the higher-order PE factor was allowed to predict the specific PH factors (M5); (e) the first-order PE dimensions were allowed to predict the specific PH factors (M6).

Currently, Mplus reports standardized coefficients separately for each level, which has been shown to be inappropriate for DL-MLSEM models for variety of reasons. Importantly, (a) contextual effects need to be properly calculated as differences between the effects obtained at both levels (Marsh et al., 2012, Morin et al., 2014), and (b) multilevel indirect effects need to be considered as they occur across levels (e.g., Preacher et al., 2010, 2011). We thus report coefficients (L1 coefficients, L2 coefficients, and contextual coefficients) that are properly standardized (β) in relation to the total (L1 and L2) variance and effect size (ES) indicators defined based on the L1 variance of the outcome using formulas provided by Marsh et al. 2009: (a) $\beta = b * SD_{predictor} / SD_{outcome}$; (b) $ES = b * SD_{predictor} / SD_{outcomeL1}$. In these formulas, b is the level-specific unstandardized regression coefficient, $SD_{predictor}$ is the level-specific standard deviation of the predictor (either L1 or L2), $SD_{outcome}$ is the combined L1 and L2 standard deviation of the outcome, and $SD_{outcomeL1}$ is the L1 standard deviation of the outcome.

Contextual effects, standardized coefficients, effect size, indirect effects, and total effects were

calculated using the multivariate delta method (Raykov & Marcoulides, 2004), implemented in Mplus using the MODEL CONSTRAINT function, where regression paths were identified with parameter labels included directly in the models and variance components were taken from the final DL-MLCFA model following procedures described by Marsh et al. (2012) and Morin et al. (2014). The annotated Mplus syntax used to estimate our a priori DL-MLSEM model, as well as the alternative models, is presented in Section 10 of the Online Supplements, whereas the commands used to obtain proper estimates of contextual effects, indirect effects[3], standardized effects, and effect sizes are presented in Section 11 of the online supplements.

Results. The results from our a priori predictive model are reported in the top section of Table 3. At L1, these results provide strong support for our a priori expectations, revealing a strong positive association between PE and PH, and a moderately strong negative relation between PH and TI. The results also show a weaker, but significant, negative relation between PE and TI. These results thus support the negative indirect association between PE and TI, as mediated by PH, while still showing additional direct negative effects of PE on TI. All of these variables are meaningful at L1 (i.e., forming contextual variables when aggregated at L2). Their L1 effects can thus all be interpreted as reflecting the fact that more highly psychologically empowered employees tend to present higher levels of psychological health and lower levels of turnover intentions and, in turn, that healthier employees also tend to present lower levels of turnover intentions. This mediation system supports well-documented associations between these constructs. Alternatively, had this model included variables reflecting inter-individual differences in perceptions of L2 climate constructs, it would have been necessary to interpret these coefficients as reflecting these inter-individual differences in perceptions, rather than the effects of true individual characteristics.

When raw L2 results are considered (which would be appropriate for climate constructs), it is first interesting to note that these results generally match their L1 counterpart, suggesting that individual effects seems to be maintained when aggregated to the group context. However, this should not come as a surprise given that these raw L2 effects simply reflect the addition of L1 effects, rather than a properly calculated contextual effect. However, these effects do appear to be slightly smaller in magnitude relative to the L1 effects. As such, these results show a moderate positive association between PE and PH, as well as a slightly smaller negative relation between PH and TI. In contrast, the direct association between PE and TI was no longer statistically significant at L2, although the indirect association between PE and TI via PH remains statistically significant. Had we been considering climate variables, these L2 coefficients resulting from the automatic group-mean centering approach would have been directly interpretable as reflecting the effects of employees shared perceptions of the group reality, with no need to go any further in terms of analyses.

Yet, when L2 contextual effects are considered, none of these associations was still statistically significant. These raw L2 results thus show that associations present at the individual level tend to generalize to the group level due to the aggregation of individual level effects, although some dilution of the magnitude of these effects is also apparent. However, the properly calculated L2 contextual effects show that neither PE nor PH seems to create either beneficial or deleterious contextual effects on employees PH and TI levels above and beyond the simple combination of individual level effects.

Finally, we considered alternative models to verify whether all observed effects could, as we expected, be summarized at the level of the higher-order PE and global PH factors (Maynard et al., 2012). The goodness-of-fit of these models is reported at the bottom Table 1. Only one of those solutions, involving associations between first-order PE factors and the specific PH factors, resulted in an improvement in model fit relative to the a priori model. The few additional statistically significant results observed in this model are reported at the bottom of Table 3. These results show that L1 ratings of competence were associated with moderately higher levels of physical job engagement beyond the

[3] When testing the statistical significance of indirect effects, it is generally recommended to rely on bootstrap confidence intervals (MacKinnon, Lockwood, & Williams, 2004), which have not yet been implemented with DL-MLSEM models. Fortunately, an alternative approach relying on Monte Carlo confidence intervals has been proposed for DL-MLSEM models (Bauer, Preacher & Gil, 2006; Preacher & Selig, 2012). Confidence intervals for all indirect effects were thus calculated with this approach (using 20,000 replications), using the online calculator proposed by Preacher and Selig (2012) and available at quantpsy.org. Detailed instructions on the use of these calculator are provided by Preacher and Selig (2012) and on the quantpsy.org, and all of the required information is provided in the "MODEL RESULTS", "TECH1" and "TECH3" sections of the Mplus output.

effects of global PE levels on PH. Likewise, L1 ratings of competence and self-determination were both associated with slightly lower levels of psychological distress beyond the effects of global PE levels on PH. L1 ratings of self-determination were also associated with moderately higher levels of morale, and both L1 and L2 levels of self-determination were associated with lower levels of disengagement over and above the effects of global PE levels on PH. However, no contextual effects of self-determination were evidenced in the results, suggesting that these L2 effects are simply the result of the aggregated L1 effects. At L2, aggregated ratings of meaningfulness also seemed to predict slightly higher levels of cognitive and emotional job engagement without, however, resulting in contextual effects. This could possibly be explained by the lower level of inter-rater reliability associated with the L2 aggregates of these two job engagement dimensions.

Benefits of DL-MLSEM

The benefits of relying on an approach allowing one to obtain multilevel estimates corrected for all forms of measurement errors (DL-MLSEM) have already been extensively documented in statistical research (e.g., Lütdke et al., 2008, 2011; Marsh et al., 2010; Morin et al., 2014), showcasing how this estimation methods provides more accurate estimate of key associations between constructs. The present study was not designed to further document these benefits, but rather to illustrate how such models should be implemented in research. However, some core advantages of the proposed methodology are readily apparent from the results obtained with our worked example.

Multilevel Measurement

From a measurement perspective, the L1 correlations observed among the PH dimensions were high enough to support the need to rely on a bifactor approach to measurement ($|r|$ = .113 to .873, M = .552). Yet, if we had only been considering measurement models estimated at a single level of analysis, alternative specifications (e.g., combining the two burnout dimensions and eliminating the psychological distress subscale) would have made it possible to obtain a more acceptable correlated factors model. However, this alternative approach would have been highly problematic from a multilevel perspective, given the much higher correlations obtained at L2 ($|r|$ = .286 to .953, M = .775), which clearly supported bifactor measurement as the only viable alternative. This apparently simple observation helps to demonstrate how measurement decisions made at an individual level of analysis cannot be expected to generalize to the group, or organizational, level of analysis.

Likewise, when contrasting the alternative measurement specification for the PE construct, measurement models estimated a single level of analysis could have been used to alternatively support a bifactor or a higher-order specification. Given the aforementioned limitation of higher-order factor models, statistically informed researchers would probably have used this information to support a bifactor operationalization. Unfortunately, this decision would also have been erroneous from a multilevel perspective, which clearly demonstrated the inadequacy of a bifactor representation of PE.

Finally, although the implications of testing, and retaining, an *isomorphic* specification of the measurement models might not have been so obvious from the present demonstration, the advantages of this specification are numerous. First, from a statistical perspective, isomorphism has been shown to result in a more stable estimation process, and in more accurate parameter estimates (Lütdke et al., 2008, 2011). In the present illustration, the former was evidenced by the fact that it was not possible to obtain convergence when estimating DL-MLCFA and DL-MLSEM models incorporating all constructs without *isomorphism*. Unfortunately, this phenomenon made it impossible to illustrate how a lack of *isomorphism* could lead to different conclusions. Second, from a more practical perspective, *isomorphism* made it possible to interpret the L2 factors in the same way as the L1 factors (Metha & Neale, 2005), which is a prerequisite condition for the estimation of contextual effects.

Climate, Context, and Centering

A core component of our illustration was focused on the illustration and how to adequately estimate and interpret climate and contextual effects. For this reason, the advantages of relying on proper methodologies in this regard should be fairly self-explanatory. However, a first key message from the present demonstration is related to centering decisions, which are often misunderstood by applied researchers due to the technical manner in which these are often presented (e.g., Enders & Tofghi, 2007), which tend to be disconnected from practical measurement and conceptual decisions. In the present article, following from Marsh et al. (2012) and Morin et al. (2014), we sought to provide more practical guidance related to the type of construct being measured. When the target of the rating is the person providing the rating, then the L2 aggregation of these ratings form a contextual construct,

the L1 component of these ratings can be directly interpreted, and the analyses require grand-mean centering. In contrast, when individual are directly asked to rate the L2 reality, then the L2 aggregation of these ratings form a climate construct, the L1 component of these ratings reflects inter-individual differences in perceptions, and the analyses require group-mean centering. As we demonstrated in the present application, centering decisions can lead to highly different conclusions regarding the L2 associations between constructs (i.e., here group-mean centered results were statistically significant, but translated into non-statistically significant grand-mean centered contextual effects).

A second key message from the present demonstration is that current estimation procedures for DL-MLSEM models automatically rely on group-mean centering irrespective of any requests for grand-mean centering made by the analyst. This is thus not simply a default estimation procedure that can be changed by the analyst. Rather, this group-mean centering is part of the mathematical underpinnings of the model. Fortunately, it is easy to convert group-mean centered results to a grand-mean centered solution by subtracting the L1 effect from its L2 counterpart (Enders & Tofighi, 2007; Marsh et al., 2012; Morin et al., 2014). We illustrated this calculation, as part of the present article, in combination with perhaps equally important calculations related to the estimation of properly standardized effects, effect size indicators, indirect effects, and total effects.

Inter-Item and Inter-Rater Reliability: An Illustration of Alternative Specifications

Our objective was to increase researchers' awareness regarding all three types of measurement errors likely to influence results in multilevel analyses, and to illustrate how to obtain estimates for all three types of measurement errors. In practice, researchers familiar with the SEM tradition tend to report, and control for, individual level estimates of inter-item reliability in their research. In contrast, researchers familiar with the MLA tradition tend to report, but not necessarily control for, multilevel estimates of inter-rater reliability in their research. In the present illustration, we argue that all three forms of reliability are important to consider, and that all of them can be controlled for as part of DL-MLSEM analyses. Our objective, however, was not to demonstrate how a lack of control for these types of errors was likely to lead to erroneous conclusions, as this has been more extensively demonstrated in the statistical research literature by Lütdke et al. (2008, 2011). Marsh et al. (2010) even demonstrated that, in combination, these three sources of unreliability could even result in phantom effects, suggesting the presence of associations that do not exist.

Alternative Specifications. Although the present data set is not ideal to illustrate the impact of alternative specifications, we still wanted to illustrate the various analytical possibilities for multilevel estimation. The first approach (Manifest-Manifest) is the one most typically used in MLA research, and relies on manifest variables (created by the summing or averaging the items forming a scale) and a manifest aggregation process (scores obtained by group members are averaged to obtain a L2 score). In a second approach (Manifest-Latent), manifest variables can be submitted to a process of latent aggregation (involving a correction for inter-rater agreement). In the present study, it is important to reinforce that a major inconvenience of relying on manifest variables lies in the impossibility to simultaneously consider global and specific facets of psychological constructs (i.e., PE and PH) in the same analysis. Thus, irrespective of the results afforded by this approach, conclusions will always be inaccurate whenever constructs are known to follow higher-order or bifactor structures.

A third and fourth alternatives involve the estimation of factors scores (FS) from preliminary single-level measurement models (which are typically estimated in organizational research), and to submit these factor scores to a process of manifest (FS-Manifest) or latent (FS-Latent) aggregation. The main advantage of factors score lies in their ability to afford some degree of control for L1 inter-item measurement error (Skrondal & Laake, 2001), as well as to preserve the L1 measurement structure (i.e., higher-order or bifactor; Morin, Boudrias et al., 2016, 2017).

A fifth alternative involves the reliance on latent variables (factors estimated from their items) at both levels, coupled with a manifest aggregation process (Latent-Manifest). A sixth alternative involves starting with a DL-MLCFA model, to save multilevel factor scores from this model, and rely on these factor scores for predictive analyses (Multilevel FS). This approach extends the aforementioned single level FS approaches by providing some degree of control for all types of measurement errors discussed in the present article, albeit this control is not as strong as that afforded by the optimal seventh alternative provided by doubly latent analyses (Doubly Latent).

Data Illustration. To illustrate how results might differ across these seven alternative approaches (Manifest-Manifest, Manifest-Latent, FS-Manifest, FS-Latent, Latent-Manifest, Multilevel FS, and

Doubly Latent[4]), we focus on the associations between the global PE factor, the global PH factor, and the TI factor (ignoring all associations involving first-order PE and specific-PH factor, which is necessary for the estimation of manifest variables). Results obtained from these seven types of models are reported in Table 3. Because all of these variables display relatively high reliability estimates, this demonstration is not optimal (i.e., TI: ω_{L1} = .828; ω_{L2} = .807; ICC2 = .749; PE: ω_{L1} = .765; ω_{L2} = .848; ICC2 = .892; PH: ω_{L1} = .844; ω_{L2} = .912; ICC2 = .769). Indeed, because of these reasonably high estimates of reliability, corrections for measurement errors might not be as critical in this study as they would be in studies including less reliable measures. However, some noteworthy differences emerge between the six partial correction models and the optimal doubly latent model. Before addressing these differences, however, we need to mention that substantial convergence difficulties were experienced in the estimation of Latent-Manifest model, which resulted in untrustworthy estimates (all close to 0). We will thus ignore the results form this model in the next paragraphs.

Level 1 Associations. A first difference is related to the size of the L1 association between PE (the construct with the lowest ω_{L1}) and PH, which is systematically underestimated in all of the partial correction models, with the possible exception of the Multilevel FS model. A similar, yet smaller, tendency can be observed for the association between PE and TI. Although, in the present application, these differences did not result in distinct conclusions (due to the strength of these associations), more meaningful differences are to be expected for smaller associations. Second, when comparing models involving or not a latent aggregation process for similar types of variables (i.e., Manifest-Manifest versus Manifest-Latent, and FS-Manifest versus FS-Latent), the L1 estimates are identical, which is aligned with the role played by latent aggregation at correcting for measurement error occurring at L2.

Raw Level 2 Associations. When considering the raw L2 effects, a similar tendency to underestimate the association between PE and PH is also observed. In addition, the association between PE and TI, although non-statistically significant across all models, is negative in most models, but positive in the FS-Latent and Doubly Latent models. This difference could reflect the lower ICC2 coefficient associated with TI, reinforcing the value of latent aggregation procedures. Finally, although the magnitude and direction of the relation between PH and TI is stable across models (although smaller in the Multilevel FS model), the statistical significance of this path (reflecting estimation accuracy, standard errors and confidence intervals) differs a lot across models: (a) $p \leq .01$ in the Manifest-Manifest and Doubly Latent models; (b) $p \leq .05$ in the Manifest-Latent and FS-Manifest model; (c) non-statistically significant in the FS-Latent and Multilevel FS models.

Contextual Effects. Contextual effects estimates show a similar tendency to underestimate the size of the association between PE and PH. Moreover, this contextual association, despite being non-statistically significant in the optimal Doubly Latent model, is estimated as statistically significant in the FS-Latent and Multilevel FS models. This difference would lead researchers to reach inaccurate conclusions regarding this contextual association. In contrast, conclusions remain essentially unchanged regarding the lack of contextual effects of PE on TI. However, this contextual effect appears to be, once again, underestimated in the Manifest-Manifest, Manifest-Latent, and FS-Manifest models. Finally, the contextual association between PH and TI shows important changes across models as it is non-statistically significant and negative in the Manifest-Manifest, FS-Manifest, and FS-Latent models, non-statistically significant in the Manifest-Latent and optimal Doubly Latent model, and positive and significant in the Multilevel FS model.

Discussion

Methodological Implications: Doubly Latent Structural Equation Models

Organizational research is inherently multilevel in nature and deserves to be considered as such. Despite this recognition, few organizational measures have been validated for multilevel assessment. Yet, the recognition that these measures reflect multilevel phenomena, and the desire to model them as such, has measurement implications that require proper statistical handling. As discussed here, any combinations of even well-validated scales that have never been used together is likely to reveal evidence of unanticipated multidimensionality, especially when brought up to the group level. Likewise, a proper handling of multilevel measurement involves recognizing the multidimensional nature of the measurement errors likely to bias results in a generally unpredictable manner and which

[4] Marsh et al. (2009) provided annotated syntax examples for the estimation of Manifest-Manifest, Manifest-Latent, Latent-Manifest, and Doubly Latent models which are easy to expand to the FS models.

require proper statistical controls. In fact, even for constructs that are typically seen as individual in nature, such as PE or PH, the recognition that influence processes involving these constructs might exist at the group or organizational level suggest that L2 variations in these constructs also need to be considered. Moreover, the climate versus context distinction reinforces the fact that even naturally individual phenomena could take up another meaning at the group level to generate a contextual influence going beyond the aggregation of the individual effects occurring among group members.

These considerations are not new. Yet, and perhaps because previous introductory efforts explicitly targeted educational (e.g., Marsh et al., 2012; Morin et al., 2014) or quantitative (Enders & Tofighi, 2007) researchers, we found little evidence of their incorporation to common practices in the organizational sciences, as was made obvious in our literature review presented in Section 2 of the online supplements focused on multilevel research on PE and PH. For this reason, we sought to provide a non-technical introduction to these important considerations for organizational researchers interested in enriching their multilevel repertoire.

Multilevel Measurement

A first issue that we highlighted is related to the importance of conducting preliminary verifications of the multilevel measurement structure of the constructs considered (e.g., Bliese et al., 2019). The possible occurrence of *jingle-jangle* confusions (i.e., construct overlap) whenever scales taken from various instruments are considered is well documented in psychological research (Casper et al., 2018; Marsh et al., 2003; Weidman et al., 2017). Less known, however, is the fact that construct overlap often tends to be higher at the group level than at the individual level (Morin et al., 2014), something that we observed in the present study. Indeed, when relying on a simple (i.e., non-hierarchical and non-bifactor) representation of PE and PH, correlations were much higher at L2 relative to L1. Additionally, clearer conclusions in terms of measurement structure can often be reached via DL-MLCFA analyses rather than via single level CFA analyses.

It is important to reinforce that empirical overlap does not mean that some measures are not necessary. Indeed, we proposed bifactor (and higher-order, pending prior theoretical and empirical support) measurement alternatives as a way to combine comprehensive assessment procedures with a way to explicitly model commonalities and specificities existing among the subscales considered (e.g., Morin et al., 2014, 2020). For this reason, bifactor models have even been previously referred to as providing a way to see both the "Forest and the Trees" in psychometric measurement (Gillet et al., 2019). In fact, even observing that relations were limited to the global PE and PH constructs should not be taken as evidence for eliminating their subscales from an assessment package. It is important to keep in mind that the global factors estimated from these models represent a synthesis of all variables incorporated within their measurement such that taking out one of those is likely to result in a change in the nature of that global construct. As shown in our secondary analyses, these models provide a way to assess the added predictive value of each subscale over that of the global factor.

Importantly, despite the fact that the measurement *isomorphism* (or equivalence) is often discussed (e.g., Bliese et al. 2007; Chan, 1998; Kozlowski & Klein, 2000) and acknowledged (e.g., Seibert et al., 2011) in the organizational sciences, we found that it was very seldom empirically verified using proper multilevel procedures and tests of metric invariance across levels. This is particularly worrisome given that *isomorphism* is required to be able to directly compare constructs across levels (Metha & Neale, 2005), to properly calculate contextual effects (Morin et al., 2014), and even to stabilize the model estimation process (Lütdke et al., 2008, 2011).

Measurement Errors

We highlighted the presence of three distinct types of measurement errors, which have the potential of resulting in biased estimates of multilevel associations among constructs (Marsh et al., 2010). Although we found evidence, in our review of multilevel PE and PH research literature, of an awareness of the first (α, ω, or ω_{L1}) and last (ICC1 and ICC2) of those measurement errors, we found a dearth of research reporting the L2 reliability of psychological constructs (ω_{L2}). Critically, despite an awareness of the role of these sources of measurement errors, typical multilevel research tends to simply assess and discuss these forms of reliability as part of preliminary analyses rather than relying on models providing a way to explicitly control their impact (D'Innocenzo et al., 2016; Huhtala et al., 2015; Kiersch, & Byrne, 2015; Xu & Yang, 2018).

Contextual and Climate Effects

A final consideration was related to the need to be as clear as possible regarding the nature of L2

constructs assessed from aggregated L1 ratings (Enders & Tofighi, 2007; Marsh et al., 2012; Morin et al., 2014). Climate constructs are formed from individual direct ratings of the L2 entity. With climate ratings, the individuals providing the ratings are not asked to provide information about themselves, but rather about the L2 reality to which they are exposed. In this situation, L1 ratings simply reflect inter-individual deviations in perceptions of the group reality, whereas L2 aggregates can be taken to reflect a more objective, or at least consensual, rating of the L2 reality. For this reason, climate constructs require group-mean centering, resulting in completely independent L1 and L2 estimates, and L1 estimates expressed as deviations from the group aggregate.

In contrast, contextual constructs are formed form the aggregation of L1 ratings reflecting an inherently L1 reality, and which may take another meaning at L2. With contextual effects, the L1 ratings (as was the case for PE and PH) are naturally meaningful and should naturally lead to matching raw L2 effects caused simply by the aggregation of L1 effects. As such, testing for contextual effects requires the estimation of a L2 relation that is controlled for, or partialled from, its L1 counterpart, a disaggregation that requires grand-mean centering (Enders & Tofighi, 2007; Marsh et al., 2012; Morin et al., 2014). Grand-mean centering thus provides a way to test if the L2 group composition has an impact on outcomes beyond the combined impact of individual characteristics.

Implication for DL-MLSEM Practice.

Our first recommendation is that any multilevel investigation in which L1 ratings are aggregated to form L2 constructs should start with an investigation of the measurement structure underlying the constructs under investigation using DL-MLCFA. This verification makes it possible to verify the adequacy of one's measurement model, to adjust this measurement model to accommodate construct overlap, and to assess metric invariance (*isomorphism*). Importantly, these preliminary measurement models also provided the information needed to assess all forms of reliability in a comprehensive manner (ω_{L1}, ω_{L2}, ICC1, and ICC2), as well as all variance estimates required for the calculation of standardized coefficients and effect sizes.

It is true that multilevel models, especially doubly latent ones, are complex and prone to estimation difficulties (nonconvergence, improper parameter estimates, etc.). For this reason, doubly latent estimation typically requires samples including a minimum of 50, but ideally more than 100, L2 units including at least 10 to 15 participants each (Lüdtke et al., 2008, 2011; Morin et al., 2014). Although the present sample met these requirements, this may not be the case for most applications. With lower sample sizes at L1 and L2, estimation difficulties are likely. Yet, we still recommend to start any multilevel investigation, even those relying on suboptimal sample sizes, by the estimation of preliminary DL-MLCFA models. Should the complexity of the model proved to be too much, then a fallback position is to start by subsets of models focusing on conceptually related constructs, as we initially did in the present study for TI, PE, and PH.

When experiencing estimation difficulties, recommendations are to rely on models including a partial correction for measurement errors, either via latent measurement or latent aggregation, but not both (Lüdtke et al., 2008, 2011; Marsh et al., 2009, 2012). The decision of which of these two corrections to incorporate should be guided by reliability assessment conducted as part of the DL-MLCFA. Researchers should primarily seek to control for the weakest form of reliability; that is, using latent aggregation procedures when ICC1 or ICC2 are low, and latent measurement procedures when ω_{L1} or ω_{L2} are low. An interesting compromise, whenever both corrections appear to be important and yet impossible to jointly implement, is to rely on factor scores from preliminary single level or multilevel measurement models to achieve a partial correction for measurement error (Skrondal & Laake, 2001). If saving multilevel factor scores proves to be impossible, then single level factor scores can still be submitted to a latent aggregation process (for a recent example, see Gagné et al., 2019). Once the best measurement model available for a specification has been selected, then multilevel predictive models can be estimated, while keeping in mind the need to rely on proper calculation for the estimation of contextual effects, standardized effects, effect sizes, indirect effects, and total effects. A more complete coverage of the various steps, decisions, and alternatives facing researchers interested to work with Dl-MLCFA and DL-MLSEM models is presented in Appendix A.

Conclusion

Doubly latent ML-SEM and ML-CFA are very powerful tools to help organizational researchers and professionals of organizational assessment measure and to study phenomena occurring at both the individual and group levels within work organizations. Despite a well-established multilevel research

tradition in the organizational sciences and a shared recognition of the critical importance to account for group-level processes in our understanding of the reality of workers and their organizational life, typical multilevel applications remain based on manifest variables (i.e., scale scores), manifest aggregation (i.e., manual aggregation of L1 ratings at L2), and unclear centering strategies. Because the organizational sciences typically rely on constructs measured with errors at both the individual, group, and individual-to-group aggregation level, this situation is unfortunate and possibly due to a lack of familiarity for modern approaches providing a solution to these limitations. Luckily, DL-MLSEM and DL-MLCFA models are now available and implemented in user-friendly statistical packages. The present research sough to introduce users to the need for, technical underpinnings, and implementation of these models, which have a broad relevance for managerial and organizational research. Because of its relative novelty and rarity, statistical research is still needed (e.g., Lüdtke et al., 2008, 2011) to define best practices. Yet, as illustrated here by our ability to address some key issues in relation to the measurement of PE and PH, as well as to their associations with TI, this statistical framework offers exciting possibilities for organizational researchers.

References

Arthur, W., Bell, S.T., & Edwards, B.D. (2007). A longitudinal examination of the comparative criterion-related validity of additive and referent-shift consensus operationalization of team efficacy. *Organizational Research Methods*, *10*, 35-58.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling*, *12*, 411-434.

Bauer, D.J.., Preacher, K.J., & Gil, K.M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models. *Psychological Methods, 11*, 142-163.

Bliese, P.D., Chan, D., & Ployhart, R. (2007). Multilevel methods: Future directions in measurement, analyses, and nonnormal outcomes. *Organizational Research Methods*, *10*, 551-563.

Bliese, P.D., Maltarich, M.A., & Hendricks, J.L. (2018). Back to basics with mixed effects models: Nine take-away points. Journal of Business and Psychology,

Bliese, P.D., Maltarich, M.A., Hendricks, J.L., Hofmann, D.A., & Adler, A.B. (2019). Improving measurement of group-level constructs by optimizing between-group differentiation. *Journal of Applied Psychology*, *104*, 293-302.

Bollen, K.C. (1989). Structural equations with latent variables. New York, NY: Wiley.

Casper, W.J., Vaziri, H., Wayne, J.H., DeHauw, S., & Greenhaus, J. (2018). The jingle-jangle of work-nonwork balance: A comprehensive and meta-analytic review of its meaning and measurement. *Journal of Applied Psychology*, *103*, 182-214.

Chan, D. (1998). Functional relations among constructs in the same content domain at different level of analysis: A typology of composition models. *Journal of Applied Psychology*, *83*, 234-246.

Chen, G., Farh, J.L., Campbell-Bush, E., Wu, Z., & Wu, X. (2013). Teams as innovative systems: Multilevel motivational antecedents of innovation in R&D teams. *Journal of Applied Psychology*, *98*, 1018-1027.

Chen, X.P., Liu, D., & Portnoy, R. (2012). A multilevel investigation of motivational cultural intelligence, organizational diversity climate, and cultural sales: Evidence from US real estate firms. *Journal of Applied psychology*, *97*, 93-106.

Chen, G., Bliese, P. D., & Mathieu, J. E. (2005). Conceptual framework and statistical procedures for delineating and testing multilevel theories of homology. *Organizational Research Methods*, *8*, 375-409.

Cheung, G.W. (2009). Introducing the latent congruence model for improving the assessment of similarity, agreement, and fit in organizational research. *Organizational Research Methods*, *12*, 6-33.

Chun, J. U., Yammarino, F. J., Dionne, S. D., Sosik, J. J., & Moon, H. K. (2009). Leadership across hierarchical levels: Multiple levels of management and multiple levels of analysis. *Leadership Quarterly*, *20*, 689–707.

Croon, M.A., & van Veldhoven, M.J.P.M. (2007). Predicting group-level outcome variables from multilevel variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*, 45-57.

D'Innocenzo, L., Luciano, M., Mathieu, J.E., Maynard, M.T., & Chen, G. (2016). Empowered to perform: A multilevel investigation of the influence of empowerment on performance in hospital units. *Academy of Management Journal*, *59*, 1290-1307.

Enders, C.K. (2010). *Applied missing data analysis*. New York, NY: Guilford.

Enders, C. K. & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models:

A new look at an old issue. *Psychological Methods, 12,* 121–138.
Fedor, D.B., Caldwell, S., & Herold, D.M. (2006). The effects of organizational changes on employee commitment: A multilevel investigation. *Personnel Psychology*, *59*, 1-29.
Gagné, M., Morin, A.J.S., Schabram, K., Wang, Z.N., Chemolli, E., & Briand, M. (2019, in press). Uncovering relations between leadership perceptions and motivation under different organizational contexts: A multilevel cross-lagged analysis. *Journal of Business & Psychology.*
Geiser, C. (2012). *Data analysis with Mplus*. New York, NY: Guilford.
Geldhof, G.J., Preacher, K.J., & Zyphur, M.J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*, 72-91.
Gignac, G.E. (2016). The higher-order model imposes a proportionality constraint: That is why the bifactor model tends to fit better. *Intelligence*, *55*, 57–68.
Gillet, N., Morin, A.J.S., Huart, I., Colombat, P., & Fouquereau, E. (2019). The forest and the trees: Investigating the globality and specificity of employees' basic need satisfaction at work. *Journal of Personality Assessment*. Early View Doi: 10.1080/00223891.2019.1591426
Gonzàlez-Romà, V., & Hernàndez, A. (2017). Multilevel modeling: Research-based lessons for substantive researchers. *Annual Review of Organizational Psychology & Organizational Behavior*, *4*, 183-210.
Gully, S.M., Incalcaterra, K.A., Joshi, A., & Beaubien, J.M. (2002). A meta-analysis of team efficacy, potency, and performance: Interdependence and level of analysis as moderators of observed relationships. *Journal of Applied Psychology*, *87*, 819-832.
Heck, R.H., & Thomas, S.L. (2015). *An introduction to multilevel modelling techniques: MLM and SEM approaches using Mplus, 3rd Edition*. New York, NY: Routledge.
Hoffman, B.J., Bynum, B.H., Piccolo, R.F., & Sutton, A.W. (2011). Person-organization value congruence: How transformational leaders influence work group effectiveness. *Academy of Management Journal*, *54*, 779-796.
Howard, J., Gagné, M., Morin, A.J.S., Wang, Z.N., & Forest, J. (2018). Using bifactor exploratory structural equation modeling to test for a continuum structure of motivation. *Journal of Management*, *44*, 2638-2664
Hox, J.J. (2010). *Multilevel analysis: Techniques and applications, 2nd Ed*. New York, NY: Routledge.
Hu, J., & Liden, R.C. (2011). Antecedents of team potency and team effectiveness: An examination of goal and process clarity and servant leadership. *Journal of Applied psychology*, *96*, 851-862.
Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.
Huhtala, M., Tolvanen, A., Mauno, S., & Feldt, T. (2015). The associations between ethical organizational culture, burnout, and engagement: A multilevel study. *Journal of Business & Psychology*, *30*, 399-414.
Ivey, G.W., Blanc, J.R.S., Michaud, K., & Dobreva-Martinova, T. (2018). A measure and model of psychological health and safety in the workplace that reflects Canada's national standard. *Canadian Journal of Administrative Sciences*, *35*, 509-522.
Kiersch, K.E., & Byrne, Z.S. (2015). Is being authentic being fair? Multilevel examination of authentic leadership, justice, and outcomes. *Journal of Leadership & Organizational Studies*, *22*, 292-303.
Klein, K.J., & Kozlowski, S.W. (2000). From micro to meso: Critical steps in conceptualizing and conducting multilevel research. *Organizational Research Methods*, *3*, 211-236.
Kline, R.B. (2016). *Principles and Practice of Structural Equation Modeling, 4th Ed*. New York: Guilford.
Laguna, M., Mielniczuk, E., & Razmus, W. (2019). Test of the bifactor model of job-related affective well-being. *Europe's Journal of Psychology*, *15*, 342-357.
Liu, D., Chen, X.P., & Yao, X. (2011). From autonomy to creativity: A multilevel investigation of the mediating role of harmonious passion. *Journal of Applied Psychology*, *96*, 294-309.
Lüdtke, O., Marsh, H.W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B.O. (2008). The Multilevel Latent Covariate model: A new, more reliable approach to group level effects in contextual studies. *Psychological Methods*, *13*, 203-229.
Lüdtke, O., Marsh, H.W., Robitzsch, A., & Trautwein, U. (2011). A 2 X 2 taxonomy of multilevel latent contextual models. *Psychological Methods*, *16*, 444-467.
MacKinnon, D.P., Lockwood, C.M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*, 99-128.
Marsh, H.W., Abduljabbar, A.S., Abu-Hilal, M., Morin, A.J.S., Abdelfattah, F., Leung, K.C., Xu, M.K., Nagengast, B., & Parker, P. (2013). Factorial, convergent, and discriminant validity of

TIMSS math and science motivation measures. *Journal of Educational Psychology, 105*, 108-128.
Marsh, H.W., Craven, R.G., Hinkley, J.W., & Debus, R.L. (2003). Evaluation of the big two factor theory of motivation orientation: Jingle-jingle fallacies. *Multivariate Behavioral Research, 38*, 189-224.
Marsh, H.W., Hau, K-T., & Grayson, D. (2005). Goodness of Fit Evaluation in Structural Equation Modeling. In A. Maydeu-Olivares & J. McArdle (Eds.), *Contemporary Psychometrics* (pp. 275-340). Hillsdale, NJ: Erlbaum.
Marsh, H.W., Hau, K.-T., Wen, Z., Nagengast, B., & Morin, A.J.S. (2013). Moderation. In T.D. Little (Ed.), *Oxford Handbook of Quantitative Methods, Vol. 2* (pp. 361-386), New York: Oxford University.
Marsh, H. W., Kuyper, H., Morin, A.J.S., Parker, P.D., & Seaton, M. (2014). Big-fish-little-pond social comparison and local dominance effects: Integrating new statistical models, methodology, design, theory and substantive implications. *Learning & Instruction, 33*, 50-66.
Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B. O., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 4*, 764–802.
Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S. & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*, 106-124.
Marsh, H.W., Scalas, L.F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg self-esteem scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment, 22*, 366-381.
Marsh, H. W., Seaton, M., Kuyper, H., Dumas; F., Huguet, P., Regner, I., Buunk, A. P., Monteil, J. M, Blanton, H., Gibbons, F. X. (2010). Phantom behavioral assimilation effects: Systematic biases in social comparison choice studies. *Journal of Personality, 78*, 671-710.
Mathieu, J.E., & Chen, G. (2011). The etiology of the multilevel paradigm in management research. *Journal of Management, 37*, 610-641.
Mathieu, J.E., Hollenbeck, J. R., van Knippenberg, D., & Ilgen, D. R. (2017). A century of work teams in the Journal of Applied Psychology. *Journal of Applied Psychology, 102*, 452-467.
Maynard, M.T., Gilson, L.L., & Mathieu, J.E. (2012). Empowerment – Fad or Fab? A multilevel review of the past two decades of research. *Journal of Management, 38*, 1231-1281,
Maynard, M.T., Mathieu, J.E., Gilson, L.L., O'Boyle, E.H., & Cigularov, K.P. (2013). Drivers and outcomes of team psychological empowerment: A meta-analytic review and model test. *Organizational Psychology Review, 3*, 101-137
McDonald, R.P. (1970). Theoretical foundations of principal factor analysis and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology, 23*, 1-21.
Mészáros, V., Ádám, S., Szabó, M., Szigeti, R., & Urbán, R. (2014). The bifactor model of the Maslach Burnout Inventory–Human Services Survey (MBI-HSS)—An alternative measurement model of burnout. *Stress & Health, 30*, 82-88.
Metha, P.D., & Neale, M.C. (2005). People are variable too: Multilevel structural equations modeling. *Psychological Methods, 10*, 259-284.
Meyer, J.P., Stanley, D.J., Herscovitch, L., & Topolnytsky, L. (2002). Affective, continuance, and normative commitment to the organization: A meta-analysis of antecedents, correlates, and consequences. *Journal of Vocational Behavior, 61*, 20-52.
Mor Barak, M.E., Nissly, J.A., & Levin, A. (2001). Antecedents to retention and turnover among child welfare, social work, and other human service employees. *Social Service Review, 75*, 625-661.
Morin, A.J.S., Boudrias, J.-S., Marsh, H.W., Madore, I., & Desrumaux, P. (2016). Further reflections on disentangling shape and level effects in person-centered analyses: An illustration exploring the dimensionality of psychological health. *Structural Equation Modeling, 23*, 438-454.
Morin, A.J.S., Boudrias, J.-S., Marsh, H.W., McInerney, D.M., Dagenais-Desmarais, V., Madore, I., & Litalien, D. (2017). Complementary variable- and person-centered approaches to exploring the dimensionality of psychometric constructs: Application to psychological wellbeing at work. *Journal of Business and Psychology, 32*, 395-419.
Morin, A.J.S., Marsh, H.W., Nagengast, B., & Scalas, L.F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *The Journal of Experimental Education, 82*, 143-167.
Morin, A.J.S., Myers, N.D., & Lee, S. (2020). Modern factor analytic techniques: Bifactor models,

exploratory structural equation modeling (ESEM) and bifactor-ESEM. In G. Tenenbaum & R.C. Eklund (Eds.), *Handbook of Sport Psychology, 4th Ed.* (pp. 1044-1073). London, UK: Wiley

Morrison, E.W., Wheeler-Smith, S.L., & Kamdar, D. (2011). Speaking up in groups: a cross-level study of group voice climate and voice. *Journal of Applied Psychology, 96*, 183-191.

Muthén, L., & Muthén, B. (2019). *Mplus user's* guide. Angeles: Muthén & Muthén.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric Theory, 3rd Edition*. New York: McGraw-Hill.

Preacher, K.J., & Selig, J.P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures, 6*, 77-98.

Preacher, K.J., Zhang, Z., & Zyphur, M.J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling, 18*, 161-182.

Preacher, K.J., Zhang, Z., & Zyphur, M.J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods, 21*, 189-205.

Preacher, K.J., Zyphur, M.J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15*, 209-233.

Quigley, N.R., Tekleab, A.G., & Tesluk, P.E. (2007). Comparing consensus- and aggregation-based methods of measuring team-level variables: The role of relationship conflict and conflict management process. *Organizational Research Methods, 10*, 589-608.

Rafferty, A.E., Jimmieson, N.L., & Armenakis, A.A. (2013). Change readiness: A multilevel review. *Journal of management, 39*, 110-135.

Raykov, T., & Marcoulides, G.A. (2004). Using the delta method for approximate interval estimation of parameter functions in SEM. *Structural Equation Modeling, 11*, 621-637.

Reise, S.P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667–696.

Rockwood, N.J. (2020). Maximum likelihood estimation of multilevel structural equation models with random slopes for latent covariates. *Psychometrika, 85*, 275-300.

Rubenstein, A.L., Eberly, M.B., Lee, T.W., & Mitchell, T.R. (2018). Surveying the forest: A meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover. *Personnel Psychology, 71*, 23-65.

Schnake, M.E., & Dumler, M.P. (2010). Levels of measurement and analysis issues in organizational citizenship behavior research. *Journal of Occupational & Organizational Psychology, 76*, 283-301.

Schneider, B., González-Romá, V., Ostroff, C., & West, M.A. (2017). Organizational climate and culture: Reflections on the history of the constructs in the Journal of Applied Psychology. *Journal of Applied Psychology, 102*, 468-482.

Seibert, S.E., Silver, S.R., & Randolph, W.A. (2004). Taking empowerment to the next level: A multiple-level model of empowerment. *Academy of Management Journal, 47*, 332-349.

Seibert, S.E., Wang, G., & Courtright, S.H. (2011). Antecedents and consequences of psychological and team empowerment in organizations. *Journal of Applied Psychology, 96*, 981-1003.

Shortell, S.M., Rousseau, D.M., Gillies, R.R., Devers, K.J.T., & Simons, T.L. (1991). Organizational assessment in intensive care units (ICUs): Construct development, reliability, and validity of the ICU Nurse-Physician Questionnaire. *Medical Care, 29*, 706-726.

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika, 66*, 563-576.

Spreitzer, G.M. (1995). Psychological empowerment in the workplace: Dimensions, measurement, and validation. *Academy of Management Journal, 38*, 1442-1465.

Tay, L., Woo, S.E., & Vermunt, J.L. (2014). A conceptual and methodological framework for psychometric isomorphism: Validation of multilevel measures. *Organizational Research Methods, 17*, 77-106.

Weidman, A., Steckler, C., & Tracy, J. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion, 17*, 267-295.

Xu, Z., & Yang, F. (2018). The cross-level effect of authentic leadership on teacher emotional exhaustion. *Journal of Pacific Rim Psychology, 12*, e35.

Zhang, Z., Zyphur, M.J., & Preacher, K.J. (2009). Testing multilevel mediation using hierarchical linear models problems and solutions. *Organizational Research Methods*, 12, 695-719.

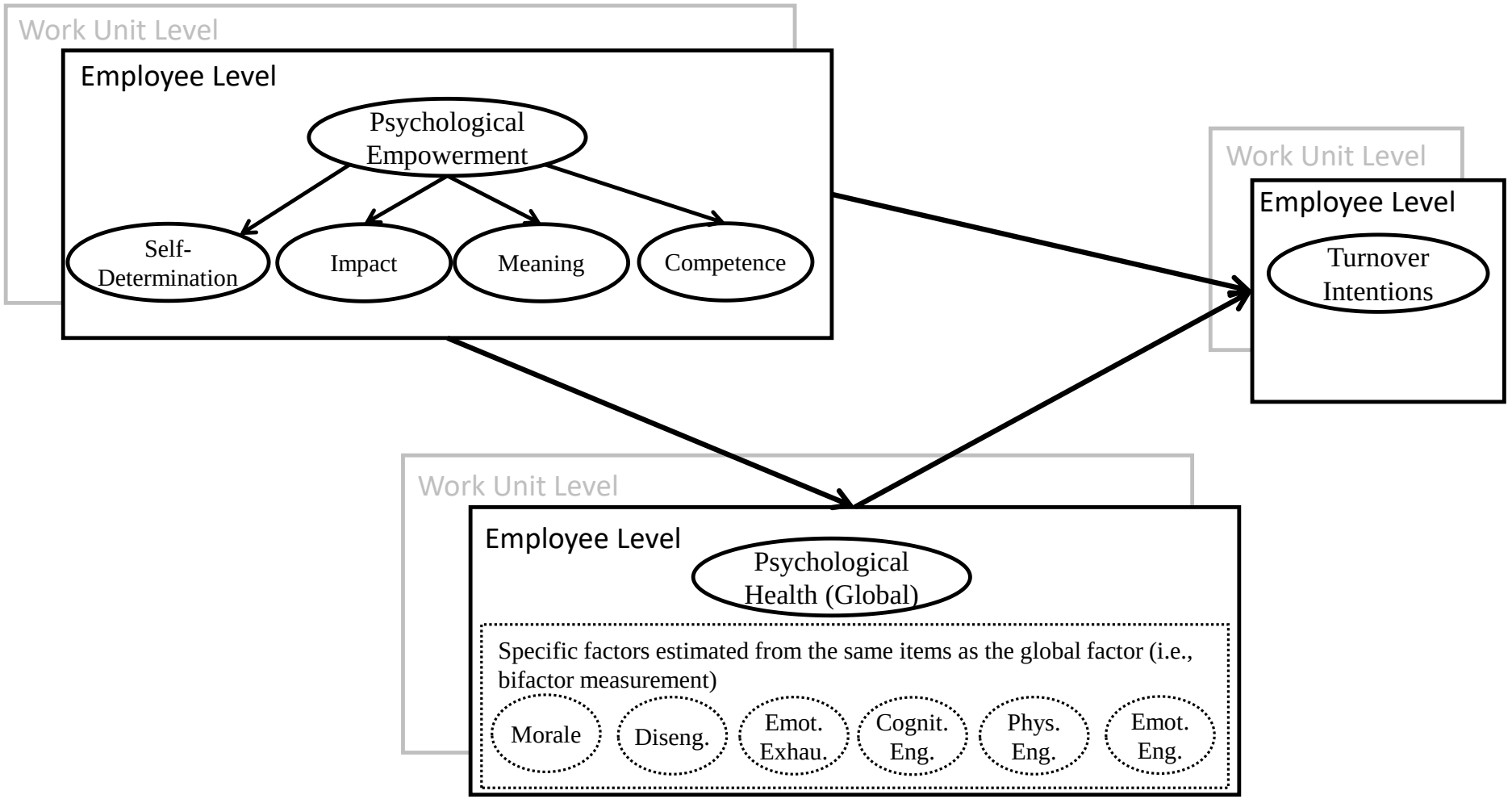

Figure 1. Theoretical model tested in the current study.

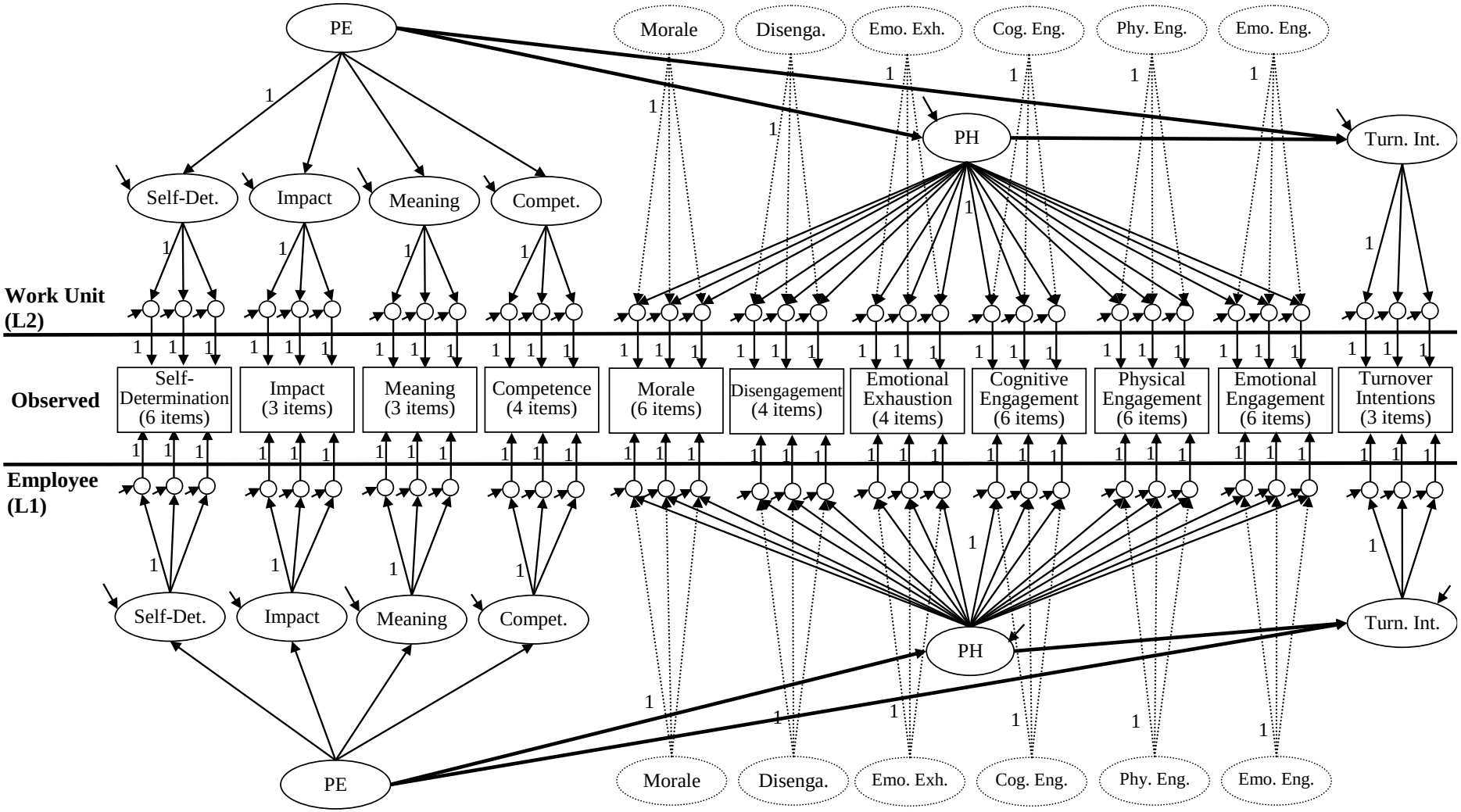

Figure 2. Doubly latent analytic model tested in the current study.

Note. PE: Psychological empowerment; PH: Psychological health; Items responses for each subscale are indicated in the rectangle appearing in the middle of the figure, and disaggregated across levels using a latent aggregation process (the small circles reflect the Level 1 and Level 2 components of these ratings); Latent factors are represented in ovals; Factor loadings linking items to the factors are represented as small black arrows (the factor loading associated with a referent indicator are fixed to one; The main predictive paths tested in the present study are represented by bold arrows; small black arrows connected to a single oval or circle reflect residual variances; to maximize the readability of the figures, factor loadings involving only three item per subscales are shown, and factor loadings associated with the specific factors (S-factors) from the bifactor measurement model used for PH are illustrated as small dotted arrows.

Table 1

Goodness-of-Fit Statistics of the Main Measurement and Predictive Models

Description	χ^2 (*df*)	CFI	TLI	RMSEA
Turnover Intentions – TI				
Multilevel CFA Invariant Factor Loadings	7.216 (2)	.998	.995	.022
Psychological Empowerment – PE (Single Level)				
Confirmatory Factor Analytic model (CFA)	814.362 (95)*	.981	.976	.036
Higher-Order CFA	819.609 (97)*	.981	.977	.036
Bifactor CFA	668.874 (85)*	.985	.978	.034
Psychological Empowerment – PE (Multilevel)				
CFA	1772.950 (193)*	.964	.955	.037
Invariant factor loadings	1539.221 (205)*	.969	.964	.033
Higher-Order CFA	1555.404 (197)*	.969	.962	.034
Invariant factor loadings	1541.696 (212)*	.970	.965	.033
Bifactor CFA	11976.412 (173)*	.735	.633	.108
Invariant factor loadings	11976.995 (200)*	.736	.683	.100
Psychological Health – PH (Single Level)				
CFA	8105.253 (796)*	.934	.929	.040
Bifactor CFA (1 global factor)	7903.357 (775)*	.936	.928	.040
Psychological Health – PH (Multilevel)				
CFA	20101.501 (1594)*	.896	.888	.044
Invariant factor loadings	15814.024 (1629)*	.920	.916	.038
Bifactor CFA (1 global factor)	16672.281 (1552)*	.915	.906	.041
Invariant factor loadings	15584.929 (1628)*	.922	.917	.038
Global Model				
Single Level CFA	13581.453 (1701)*	.929	.924	.034
Multilevel CFA (Invariant factor loadings)	27919.460 (3500)*	.913	.909	.034
Predictive Models				
M1. Multilevel Predictive: A priori model	28103.465 (3528)*	.913	.909	.034
M2. M1 + Specific PE —> TI	28251.660 (3520)*	.912	.909	.035
M3. M1 + Specific PH —> TI	27949.629 (3514)*	.913	.909	.034
M4. M1 + Specific PE —> Global PH	27777.840 (3520)*	.913	.909	.034
M5. M1 + Global PE —> Specific PH	32167.964 (3500)*	.898	.893	.037
M6. M1 + Specific PE —> Specific PH	23772.239 (3458)*	.928	.924	.032

Note. * $p < .01$; χ^2: robust chi-square test of exact fit; *df*: degrees of freedom; CFI: comparative fit index; TLI: Tucker-Lewis index; RMSEA: root mean square error of approximation; 90% CI: 90% confidence interval.

Table 2

Reliability Information

	ω	ω_{L1}	ω_{L2}	ICC1	ICC2
Turnover Intentions	.827	.828	.807	.024	.749
Psychological Empowerment					
Self-Determination (First-Order Factor)	.797	.786	.935	.077	.909
Impact (First-Order Factor)	.920	.917	.950	.044	.847
Meaning (First-Order Factor)	.963	.962	.998	.051	.865
Competence (First-Order Factor)	.892	.889	.955	.037	.822
Empowerment (Higher-Order Factor)	.769	.765	.848	.064	.892
Psychological Health					
Global Psychological Health (G-Factor)	.977	.976	.995	.065	.892
Morale (S-Factor)	.855	.852	.923	.029	.782
Physical Engagement (S-Factor)	.915	.915	.907	.012	.600
Emotional Engagement (S-Factor)	.866	.864	.910	.014	.638
Cognitive Engagement (S-Factor)	.928	.928	.899	.006	.420
Emotional Exhaustion (S-Factor)	.514	.517	.474	.019	.702
Disengagement (S-Factor)	.437	.419	.788	.077	.909
Psychological Distress (S-Factor)	.846	.844	.912	.027	.769

Note. G: Global factor from a bifactor measurement model; S: Specific factor from a bifactor measurement model; ω: omega coefficient of model-based composite reliability estimated from a single level model; ω_{L1}: omega coefficient of model-based composite reliability estimated at level 1 (individual) from a multilevel model; ω_{L2}: omega coefficient of model-based composite reliability estimated at level 2 (unit) from a multilevel model; ICC1: Intraclass correlation; ICC2: Reliability of unit-level aggregations.

Table 3

Multilevel Predictive Results

	Raw Multilevel Effects			Contextual Effects		
	Est. [CI]	Std. [CI]	ES [CI]	Est. [CI]	Std. [CI]	ES [CI]
Level 1						
PE —> PH	2.417 [2.072; 2.761]**	.888 [.761; 1.014]**	.918 [.787; 1.049]**			
PE —> TI	-.238 [-.409; -.067]**	-.094 [-.162; -.027]**	-.096 [-.164; -.027]**			
PH —> TI	-.615 [-.679; -.552]**	-.658 [-.726; -.590]**	-.666 [-.735; -.608]**			
Indirect Effect	-1.487 [-1.740; -1.248]**					
Total Effect	-1.725 [-1.983; -1.468]**					
Level 2						
PE —> PH	2.760 [2.212; 3.307]**	.269 [.216; .322]**	.278 [.223; .333]**	.343 [-.055; .741]	.033 [-.005; .072]	.035 [-.006; .075]
PE —> TI	.117 [-.802; 1.037]	.012 [-.084; .109]	.012 [-.085; .110]	.356 [-.590; 1.301]	.037 [-.062; .137]	.038 [-.063; .139]
PH —> TI	-.500 [-.803; -.197]**	-.141 [-.227; -.056]**	-.143 [-.230; -.056]**	.115 [-.194; .425]	.032 [-.055; .120]	.033 [-.056; .121]
Indirect Effect	-1.380 [-2.274; -.535]**			.040 [-.060; .223]		
Total Effect	-1.263 [-1.642; -.884]**			.395 [-.443; 1.233]		
	Additional Raw Effects			Additional Contextual Effects		
	Est. [CI]	Std. [CI]	ES [CI]	Est. [CI]	Std. [CI]	ES [CI]
Level 1						
PES —> Mor.	.979 [.274; 1.684]**	.342 [.096; .588]**	.348 [.097; .599]**			
PES —> Dis.	-1.321 [-1.967; -.675]**	-1.466 [-2.183; -.749]**	-1.525 [-2.272; -.779]**			
PES —> PD	-.243 [-.461; -.024]*	-.200 [-.381; -.020]*	-.202 [-.385; -.020]*			
PEC—> PD	-.012 [-.022; -.001]*	-.039 [-.075; -.004]*	-.040 [-.076; -.004]*			
PEC —> JEP	.191 [.026; .357]*	.199 [.027; .372]*	.201 [.027; .374]*			
Level 2						
PEM —> JEE	.384 [.074; .695]*	.074 [.014; .133]*	.074 [.014; .134]*	.095 [-.390; .579]	.018 [-.075; .111]	.018 [-.075; .112]*
PEM —> JEC	.538 [.050; 1.026]*	.093 [.009; .177]*	.094 [.009; .178]*	.284 [-.698; 1.266]	.049 [-.121; .219]	.049 [-.121; .220]
PES —> Dis.	-.945 [-1.542; -.348]**	-.185 [-.302; -.068]*	-.193 [-.315; -.071]*	.376 [-.068; .821]	.074 [-.013; .161]	.074 [-.014; .168]

Note. * = $p \leq .05$; ** = $p \leq .01$; PE: global psychological empowerment (PES: Self-Determination, PEC: Competence; PEM: Meaning); PH: global psychological health [Mor.: Morale; Dis.: Disengagement; JE: Job engagement (JEP: Physical job engagement; JEE: Emotional job engagement; JEC: Cognitive job engagement)]; TI: turnover intentions; Est. = unstandardized estimate; Std. = standardized estimate; ES: effect size; CI: 95% confidence interval; confidence intervals for the coefficients themselves, as well as for the total effects, are parametric confidence intervals (i.e., estimate +/- 1.96*standard error; confidence intervals for all indirect effects are nonparametric Monte Carlo confidence intervals (Preacher, & Selig, 2012)

Table 4

Results from Alternative Model Specifications

	Manifest Variables Manifest Aggregation Est. [CI]	 Latent Aggregation Est. [CI]	Single Level FS Manifest Aggregation Est. [CI]	 Latent Aggregation Est. [CI]	Latent Variables Manifest Aggregation Est. [CI]	Multilevel FS Est. [CI]	Doubly Latent (From Table 3) Est. [CI]
Level 1							
PE —> PH	.669 [.646; .693]**	.670 [.646; .693]**	.897 [.884; .911]**	.897 [.884; .911]**	.082 [-.114; .277]	1.853 [1.785; 1.920]**	2.417 [2.072; 2.761]**
PE —> TI	-.149 [-.182; -.116]**	-.149 [-.182; -.116]**	-.268 [-.308; -.228]**	-.268 [-.308; -.227]**	-.332 [-1.696; 1.033]	-.406 [-.484; -.328]**	-.238 [-.409; -.067]**
PH —> TI	-.545 [-.579; -.510]**	-.545 [-.579; -.510]**	-.593 [-.636; -.550]**	-.593 [-.636; -.550]**	.061 [-.892; 1.013]	-.481 [-.510; -.452]**	-.615 [-.679; -.552]**
Level 2 (Raw)							
PE —> PH	.763 [.608; .917]**	.731 [.569; .893]**	.968 [.905; 1.031]**	.968 [.904; 1.032]**	.000 [-1.032; 1.031]	2.381 [1.896; 2.867]**	2.760 [2.212; 3.307]**
PE —> TI	-.015 [-.483; .453]	-.017 [-.402; .367]	-.138 [-.868; .592]	.092 [-.905; 1.089]	-.008 [-7.894; 7.808]	-.185 [-.934; .565]	.117 [-.802; 1.037]
PH —> TI	-.703 [-1.212; -.194]**	-.492 [-.939; -.045]*	-.795 [-1.533; -.058]*	-.791 [-1.796; .214]	.002 [-1.404; 1.407]	-.218 [-.461; .025]	-.500 [-.803; -.197]**
Level 2 (Contextual)							
PE —> PH	.093 [-.062; .248]	.062 [-.062; .248]	.071 [.007; .135]*	.070 [.007; .135]*	-.082 [-1.055; .890]	.529 [.032; .1.026]*	.343 [-.055; .741]
PE —> TI	.134 [-.335; .603]	.132 [-.335; .603]	.129 [-.602; .861]	.359 [-.602; .861]	.324 [-6.475; 7.123]	.222 [-.522; .965]	.356 [-.590; 1.301]
PH —> TI	-.158 [-.667; .350]	.053 [-.667; .350]	-.202 [-.942; .538]	-.198 [-.942; .538]	-.059 [-1.597; 1.479]	.264 [.022; .506]*	.115 [-.194; .425]

Note. * = $p \leq .05$; ** = $p \leq .01$; FS: Factor score; PE: global psychological empowerment; PH: global psychological health; TI: turnover intentions; Est. = unstandardized estimate; CI: 95% confidence interval (parametric).

Appendix A

A Quick Guide to Doubly Latent Estimation

Step 1. Data Collection

[1] Sample size: Doubly latent estimation requires samples including 50, but ideally over 100, L2 units, each formed of at least 10 to 15 participants (Lüdtke et al., 2008, 2011; Morin et al., 2014). With smaller samples, partial correction models might be required (see Step 4).

[2] Measurement: Psychometrically validated measures with a clear referent (either the person or the work group, rather than a mixture of both) should be used. For climate constructs, the referent should be the work group. For contextual constructs, the referent should be the individual. Other measures can also be integrated for use at a single level of analysis, and these should also have a clear referent.

Step 2: Preliminary Measurement Models

[3] Correlated Factors (CFA) Models: Using single level correlated factors (CFA) measurement models, the *a priori* measurement structure of all variables should be assessed following generally accepted guidelines for the estimation of CFA/SEM models (e.g., Kline, 2016; Geiser, 2012). This step makes it possible to locate obvious measurement problems prior to the estimation of more complex doubly latent models.

[4] Bifactor Models. Bifactor models should be estimated and contrasted with correlated factors models whenever: (a) there are theoretical or empirical reasons to do so; (b) correlations estimated as part of the correlated factors model are high enough (e.g., $r \geq .700$ or .800) to suggest conceptual redundancies. Bifactor models should be favored relative to correlated factors models when: (a) the correlated factors model reveal high factor correlations; (b) the bifactor model result in an equivalent or higher level of fit to the data; (b) the bifactor model result in a well-defined G-factor accompanied by at least some well-defined S-factors (i.e., satisfactory factor loadings and a level of composite reliability ideally $\omega \geq .500$; Perreira et al., 2018; Morin et al., 2020).

[5] Higher-Order Models. Higher-order models should only be estimated when there are strong theoretical or empirical reasons supporting their use. When estimated, they should always be contrasted with bifactor models, and they should only be retained when their level of model fit is equivalent or superior to their bifactor counterparts.

[6] Alternatives. When the correlated factors model reveal high factor correlations but bifactor models do not make sense (i.e., when it does not make sense to estimate a global factor), then alternatives are to: (a) combine the variables that are too highly correlated; (b) delete one of the variables that are too highly correlated; (c) rely on a partial bifactor model (including the subset of the variables that can be grouped logically).

[7] Troubleshooting.

- *Misfit.* When none of the alternative models is able to achieve a proper level of fit to the data, then the analyst should consider revising this model based on troubleshooting guidelines typically used on CFA/SEM (e.g., Kline, 2016): Constructs can be eliminated, factors can be combined or separated, modification indices can be examined, etc.
- *Convergence*: When the models fail to converge, iterations can be increased (e.g., ITERATIONS = 10000; H1ITERATIONS = 10000; MITERATIONS = 10000). When this is not sufficient, convergence criteria can be decreased to .0001 (CONVERGENCE = .0001; H1CONVERGENCE = .0001; MCONVERGENCE = .0001), then to .0005, .001, .005, .01, and .05. When this is not sufficient to achieve convergence, then models should be estimated using subsets of conceptually-related variables (as we did) or variables should be eliminated.
- *Negative variances estimates*: Improper parameter estimates (e.g., the negative variance or residual estimates) can be labelled using an alphanumeric expression in parentheses [e.g., Var1 (cons1);] and this label should be used in the MODEL constraint section of the syntax to force the estimation of a proper solution MODEL CONSTRAINT: cons1 > 0;). For more information on this approach often required in doubly latent estimation procedures, see Marsh et al. (2012) and Morin et al. (2014).

Step 3: Multilevel Measurement Models

[8] Model Estimation and Comparisons: All of the alternative models found to be plausible (i.e., able to achieve a satisfactory level of model fit) in Step 2 should be converted to DL-MLCFA models. An estimation and decision process identical to that highlighted in Step 2 can be applied at this step, and similar troubleshooting approaches can be used. As in Step 2, this estimation process should, ideally, be conducted while including all constructs relevant to the study's objectives. When this is not possible, subsets of models should be estimated.

[9] Isomorphism:

- When constructs are assessed, and relevant, at a single level of analysis, *isomorphism* does not have to be tested.
- When different sets of matching items (i.e., using a similar number of items with a similar content but differing only in their referent) are used to assess similar constructs (i.e., constructs with the same number of dimensions sharing a logically related signification), *isomorphism* can be tested when: (a) there are theoretical or empirical reasons to expect these constructs to have an isomorphic structure across levels; (b) when there are no theoretical or empirical counterindications in order to increase the parsimony and stability of estimation of the model. Here, failure to support isomorphism should only be considered as a result of the analyses, which can be pursued without isomorphism.
- When doubly latent constructs are assessed (when the same ratings are separated into individual and group components using latent aggregation), then isomorphism has to be tested. For contextual constructs, *isomorphism* needs to be supported by the data (i.e., resulting in a similar, better, or not drastically reduced level of model fit relative to the non-*isomorphic* model). In this context, failure to obtain *isomorphism* should lead to a complete re-assessment of the study's objectives and/or to the decision to exclude this construct from the model. Otherwise, alternative measurement specifications (i.e., deleting items and/or dimensions) could be investigated until an *isomorphic* model can be located. For climate constructs, *isomorphism* also needs to be supported. However, failure to obtain support for isomorphism simply indicate that the L1 counterparts of the L2 climate effects should not be interpreted substantively (but they can still be controlled for as part of the model).

[10] Final Model: A final measurement model including all constructs should be estimated. Often, when the estimation of such an integrative model has proven impossible in the previous stages, the final *isomorphic* operationalization of all constructs can still be estimated in a single model (due to its greater parsimony). This is what happened in the present study. In some situations, however, it will remain impossible to estimate a doubly latent multilevel model including all constructs. In these situations, it will also be impossible to estimate the whole predictive model in a single step. For this reason, partial correction models (see Step 4) can be estimated, or factor scores can be extracted for a subset of constructs to maintain some degree of control for measurement errors. These factor scores should ideally be taken from multilevel measurement models (i.e., affording a partial correction for all three types of error) but can also be taken from single level measurement models in extreme situations (i.e., affording a partial correction for L1 inter-item measurement errors and making it possible to rely on a latent aggregation process.

Step 4: Measurement Errors

[11] Reliability Assessment. The final DL-MLCFA integrative model or the final subsets of DL-MLCFA models should be used to assess: (a) inter-item reliability at the individual level (ω_{L1}); inter-item reliability at the group level (ω_{L2}); (c) inter-rater agreement in collective ratings (ICC2); (d) the proportion of variance occurring at L1 and L2 (ICC1).

- *L2 variability.* ICC1 values should be ideally greater than .100 but minimally greater than .050 to justify relying on multilevel procedures (e.g., Gonzàlez-Romà, & Hernàndez, 2017; Lütdke et al., 2008, 2011). In some situations (and this happened here), some constructs will have an ICC1 lower than .050. In the context of bifactor or higher-order models, this guideline is particularly important for the main construct(s) under-investigation, which is typically the G-factor (bifactor) or the higher-order factor (higher-order models). Furthermore, this guideline is also more important for exogenous (predictors, with no

predictors of their own) constructs than for endogenous (outcomes) ones. Indeed, even when the outcome of a multilevel regression model is a pure individual level construct, any attempt to investigate the role of group-level predictors in relation to this construct will need to be estimated at level 2 (e.g., Zhang et al., 2009). Thus, for secondary factors (e.g., S-factors and first-order factors) and outcomes, observing a lower than ideal ICC1 suggests that interpretations of the results associated with the L2 component of these constructs should be taken with a grain of salt. Thus, even in the presence of relative strong L2 associations, the analyst should keep in mind that very little variance occurs at this level. In contrast, observing a lower than ideal ICC1 in relation to the main constructs, especially for exogenous predictors, should lead the analyst to re-assess the appropriateness of the model.

- *Low reliability.* As in any investigation, reliability estimates (ω_{L1}, ω_{L2}, ICC2) should generally be higher than .6, ideally .7. However, for the S-factors estimated within bifactor models, because true score variance is divided into two different sets of factors (i.e., the G-factor and the S-factors), lower reliability estimates (e.g., .5) remain acceptable (Perreira et al., 2018; Morin et al., 2020). Despite this generic interpretation guideline, doubly latent models include a natural correction for these types of measurement errors, which means that observing lower levels of reliability should not be considered to be a problem, as long as the proper doubly latent estimation procedures can be applied. Beyond this, however, results associated with constructs with very low reliability estimates (e.g., lower than .400) should be interpreted with a great deal of caution, as these very low estimates do not only call into question the reliability of these constructs, but rather suggest that these constructs might not be measured properly and may mainly reflect random noise.

[12] Partial correction: In addition to providing useful information about the constructs included in the study, these various estimates of reliability are particularly important to consider when doubly latent estimation is impossible (when the theoretical model is too complex in relation to the available sample). For these situations, we highlighted in the manuscript a variety of partial correction models (Manifest-Manifest, Manifest-Latent, FS-Manifest, FS-Latent, Latent-Manifest, Multilevel FS). The decision of which of these approach to retain should be, first and foremost, guided by the results from the analyses of reliability conducted in this step. More precisely, latent aggregation should be favored when ICC2 are low, whereas latent measurement (or factor scores) should be favored when estimates of composite reliability (ω_{L1}, ω_{L2}) are low.

Step 5: Predictive Models

[13] Predictive model. Once a satisfactory final measurement model has been estimated, then this model can be converted to the researcher's a priori predictive model. Alternatively, once the final sets of measurement models have been estimated and the desired partial correction approach has been identified, the resulting predictive model can be estimated. As in any applications of SEM or MLA, alternative specifications can be compared based on model fit information and an examination of parameter estimates (as illustrated in this article).

[14] Centering:

- The effects of constructs estimated only at L2 should be estimated via grand-mean centering.
- The effects of constructs estimated only at L1 and representing individual characteristics should be estimated while requesting group-mean centering.
- The effects of constructs estimated only at L1 and conceptualized as inter-individual variations around a group average should be estimated while requesting grand-mean centering.
- The effects of doubly latent climate constructs will be automatically estimated using proper group-mean centering.
- The effects of doubly latent contextual constructs will be automatically estimated using improper group-mean centering. These effects will thus need to be calculated via the procedures outlined in section 11 of the online supplements (i.e., the L1 effect would need to be subtracted from its L2 counterpart).

[15] Standardized effects. Standardized effects provided automatically by Mplus are improper. Properly standardized effects, effect size indicators, and indirect effects should be calculated using the procedures outlined in section 11 of the online supplements.

[16] S-factors. It is important to keep in mind that the effects associated with S-factors obtained from a bifactor model differ greatly from that of their first-order counterparts. Indeed, rather than reflecting the variance shard among the subset of items forming this measure, S-factors reflect the variance that is shared among these indicators once that explained by the G-factor has been extracted. As such, they should be interpreted as reflecting deviations on these specific constructs relative to each person score on the global construct (on the G-factor).

Online Supplements for:

Doubly Latent Multilevel Procedures for Organizational Assessment and Prediction

Section 1

Description of the Measures used in the Worked Example

As noted in the main manuscript, the majority of the respondents (76%) completed the English version of the questionnaires, whereas the remaining completed the French version. For the few measures (i.e., job engagement, morale and turnover intentions) not already validated in both official languages of Canada (English and French), professional translators from the Government of Canada's Translation Bureau translated the original English items into French. Bilingual experts then back-translated these items into English. Discrepancies were discussed and resolved by consensus. Given that the sample includes a mixture of male and female civilian and military respondents who completed English or French questionnaires, tests of measurement invariance of individual responses as a function of these three categorization of respondents (male/female, military-civilian, English/French) were also realized (Cheung, 2008; Millsap, 2011). As shown in Table S4 of the online supplement, full invariance was observed for each categorization.

Psychological Empowerment. Feelings of meaning (3 items; α = .97; e.g., *The work I do is meaningful to me*) and impact (3 items; α = .92; e.g., *I have significant influence over what happens in my department*) were assessed with the relevant subscales from Spreitzer's (1995; French version by Boudrias, Rousseau, Migneault, Morin, & Courcy, 2010) questionnaire. Feelings of self-determination (6 items; α = .81; e.g., *I feel free to do my job the way I think it could best be done*) and competence (4 items; α = .81; e.g., *I am good at the things I do in my job*) were assessed with the relevant subscales from Van den Broeck, Vansteenkiste, De Witte, Soenens, and Lens' (2010; French version by Gillet, Morin, Huart, Colombat, and Fouquereau, 2019) questionnaire. All items were rated on a 5-point response scale ranging from 1-*Totally Disagree* to 5-*Totally Agree*.

Burnout. Participants' levels of disengagement (4 items; α = .82; e.g., *Over time, one can become disconnected from this type of work*) and emotional exhaustion (4 items; α = .86; e.g., *During my work, I often feel emotionally drained*) were measured with an 8-item short form of the Oldenburg Burnout Inventory (Demerouti, Bakker, Vardakou, & Kantas, 2003; French version by Chevrier, 2009). All items were rated on a 4-point scale ranging from 1-*Strongly Disagree* to 4-*Strongly Agree*.

Job Engagement. Cognitive (6 items; α = .95; e.g., *At work, I am absorbed by my job*), physical (6 items; α = .93; e.g., *I work with intensity on my job*), and emotional (6 items; α = .95; e.g., *I am proud of my job*) engagement were assessed with Rich, Lepine, and Crawford's (2010) measure. All items were rated on a 5-point scale ranging from 1-*Strongly Disagree* to 5-*Strongly Agree*.

Morale. Morale was assessed with the six-item (α = .94; e.g., *Your level of drive*) Military Morale Scale (Britt & Dickinson, 2006). Participants rated each item on a five-point scale ranging from 1-*Very Low* to 5-*Very High*, as a function of the stem "*The following items refer to your motivation and enthusiasm for accomplishing work objectives. Please rate the following items using the scale below*".

Psychological Distress. The 10-item (α = .93; e.g., *did you feel depressed?*) Kessler Psychological Distress Scale (Kessler et al., 2002; French version by Arnaud et al., 2010) was used to assess this construct. All items were rated on a 5-point frequency scale ranging from 1-*None of the Time* to 5-*All the Time* and presented after the stem "*In the last four weeks, about how often…".*

Turnover Intentions. Participants' intentions to leave their current employment were measured using Colarelli's (2984) three-item (α = .82; e.g., *I frequently think of quitting my job*) measure. All items were rated on a five-point scale ranging from 1-*Strongly Disagree* to 5-*Strongly Agree*.

References

Arnaud, B., Malet, L., Teissedre, F., Izaute, M., Moustafa, F., Geneste, J., Schmidt, J., Llorca, P., Brousse, G. (2010). Validity study of Kessler's psychological distress scales conducted among patients admitted to French emergency department for alcohol consumption-related disorders. *Alcoholism: Clinical and Experimental Research, 34*, 1235-1245.

Boudrias, J.-S., Rousseau, V., Migneault, P., Morin, A.J.S., & Courcy, F. (2010). Habilitation psychologique: Validation d'une mesure en langue Française. [Psychological empowerment: a French measure]. *Swiss Journal of Psychology, 69*, 147-159.

Britt, T.W., & Dickinson, J.M. (2006). Morale during military operations: A positive psychology approach. In T.W. Britt, C.A. Castro, & A.B. Adler, (Eds.), *Military life: The psychology of serving in peace and combat* (pp. 157–184). Westport, CT: Praeger Security Int.

Cheung, G.W. (2008). Testing equivalence in the structure, means, and variances of higher-order constructs with structural equation modeling. *Organizational Research Methods*, *11*, 593-613.

Chevrier, N. (2009). Adaptation Québécoise de l'Oldenburg Burnout Inventory (OLBI) [Quebec validation of the Oldenburg Burnout Inventory (OLBI)]. Unpublished doctoral thesis, Montreal, Canada: Université du Québec a Montréal

Colarelli, S.M. (1984). Methods of communication and mediating processes in realistic job previews. *Journal of Applied Psychology, 69*, 633–642.

Demerouti, E., Bakker, A.B., Vardakou, I., & Kantas, A. (2003). The convergent validity of two burnout instruments: A multitrait-multimethod analysis. *European Journal of Psychological Assessment, 19*, 12–23.

Gillet, N., Morin, A.J.S., Huart, I., Colombat, P., & Fouquereau, E. (2019). The forest and the trees: Investigating the globality and specificity of employees' basic need satisfaction at work. *Journal of Personality Assessment*. Early View Doi: 10.1080/00223891.2019.1591426

Kessler, R.C., Andrews, G., Colpe, L.J., Hiripi, E., Mroczek, D.K., Normand, S.-L.T., Walters, E.E., Zaslavsky, A.M. (2002). Short screening scales to monitor population prevalence and trends in non-specific psychological distress. *Psychological Medicine, 32*, 959–976.

Millsap, R.E. (2011). *Statistical approaches to measurement invariance*. New York: Taylor & Francis.

Rich, B.L., Lepine, J.A., & Crawford, E.R. (2010). Job engagement: Antecedents and effects on job performance. *Academy of Management Journal, 53*, 617–635.

Spreitzer, G.M. (1995). Psychological empowerment in the workplace: Dimensions, measurement, and validation. *Academy of Management Journal*, *38*, 1442-1465.

Van den Broeck, A., Vansteenkiste, M., De Witte, H., Soenens, B., & Lens, W. (2010). Capturing autonomy, competence, and relatedness at work. *Journal of Occupational & Organizational Psychology*, *83*, 981-1002.

Section 2

Theoretical Introduction to the Worked Example

Decades of research in several areas of the social sciences have underscored the importance of psychological health (PH) to organizations (e.g., Day & Randell, 2014; Harter, Schmidt, & Keyes, 2003). In Canada alone, the purely economic costs of low workplace PH reached $10-$15 billion a year in 2006 (Baynton & Fournier, 2017). This recognition has led the Mental Health Commission of Canada (MHCC) to commission a national standard for psychological health and safety in the workplace (Canadian Standards Association Group and Bureau de normalisation du Québec [CSA/BNQ], 2013). Since its launch, a variety of Canadian organizations have shown interest in its adoption. In particular, the publication of these standards led the CAF/DND to develop an organizational assessment test battery, called the Unit Morale Profile version 2.0 (UMP2) in order to conduct comprehensive assessments of PH and its determinants among CAF/DND personnel (Ivey, Michaud, Blanc, & Dobreva-Martinova, 2018). The analyses presented in the main manuscript provide a validation of the components of this framework that are related to the direct assessment of PH. The PH components integrated to this comprehensive assessment package include two components of burnout (emotional exhaustion and disengagement) taken from the Oldenburg Burnout Inventory (Demerouti, Bakker, Vardakou, & Kantas, 2003), three components of job engagement (cognitive, emotional, and physical) taken from the Job Engagement Scale (Rich, Lepine, & Crawford, 2010), the Military Morale scale (Britt & Dickinson, 2006), and well as the Psychological Distress Scale (Kessler et al., 2002). As is typical in organizational assessments, CAF/DND made no claim that these components would cover the whole range of variables that could be considered to be part of employees' PH at work. Rather, these variables were selected as those that were the most appropriate to consider in terms of being both aligned with the new Canadian standard and relevant to the Canadian Defence Team context.

In addition, the analyses presented in the main manuscript also consider the role of psychological empowerment (PE; Spreitzer, 1995) as a key individual resource involved in the prediction of PH, and employees' turnover intentions as an outcome of both PE and PH. More importantly, these analyses also investigate whether these associations translate to the work unit level (L2), and whether PE and PH carry contextual effects at these levels over and above what could be expected from the aggregation of their individual effects.

Turnover Intentions

The interest of focusing on turnover intentions as the distal outcome variable stems from four different reasons. First, turnover intensions are known to represent the main predictor of voluntary turnover for a variety of organizations (Heavey, Holwerda, & Hausknecht, 2013; Rubenstein, Eberly, Lee, & Mitchell, 2018), an association reported to be particularly important for military organizations (Griffeth, Hom, & Gaertner, 2000; Lytell, & Drasgow, 2009). Second, the cost of turnover intentions and turnover in terms of organizational performance, and recruitment and training are known to be immense (Hancock, Allen, Bosco, McDaniel, & Pierce, 2013; Heavey et al., 2013; Park & Shaw, 2013). Third, meta-analytic reviews focused on the antecedents of turnover intentions (Mor Barak, Nissly, & Levin, 2001; Rubenstein et al., 2018), meta-analytic reviews on the consequences of PE (Maynard et al., 2012; Seibert, Wang, & Courtright, 2011), or studies focused on the outcomes of various indicators of PH (e.g., Brunetto, Teo, Shaklock, & Farr-Wharton, 2012; Wu, Rafiq, & Chin, 2017; Yanchus, Periard, & Osatuke, 2017) all converge on the presence of well-established relations, at the individual level, between both PE and PH, and turnover intentions. The fact that these relations are so well established at the individual level (L1), while remaining understudied at the group level (L2), makes turnover intentions an ideal outcome for the assessment of whether L1 effects would translate to L2 and whether these effects would be contextual in nature. Finally, despite repeated calls for the adoption of a multilevel approach to the study of turnover intentions (Hom, Lee, Shaw, & Hausknecht, 2017; Rubenstein et al., 2018), true multilevel research in the determinants of turnover intentions remains scarce.

PE as a Core Psychological Resource for PH

PE refers to a "*set of psychological states that are necessary for an individual to feel a sense of control in relation to their work*" (Spreitzer, 2008, p. 56) and that form a core psychological resource

for employees, allowing them to play a volitional role at work (Spreitzer, 1995; Thomas, & Velthouse, 1990). PE encompasses the four work-related cognitions of competence, self-determination, impact, and meaning. *Competence* refers to feelings of having the abilities required for a successful execution of one's work, a cognition close to the concept of self-efficacy. *Self-determination* refers to feelings of being in control when initiating and regulating work-related behaviors, and is akin to the concept of autonomy. *Impact* refers to feelings of being able to influence operational, strategic, or administrative outcomes at work. Finally, *meaning* refers to feelings that there is a good fit between the requirement of their work and employees' own personal beliefs, standards and values. Despite their distinct nature, these four cognitions have been systematically shown to follow a higher-order representation according to which they all reflect an overarching *isomorphic* PE construct at the individual and group (i.e., team, unit) levels (Morin, Meyer et al., 2016; Seibert et al., 2011).

PE value for military organizations (Rasmussen, 2012) stems from its role as a psychological resource that can help workers handle the inherent stressfulness of their work (Raymer, Lindsay, & Watola, 2017) and counterweights the structured hierarchical reality of military life (Campbell & Campbell, 2011; Kotze, Menon, & Vos, 2007). Meta-analyses support the role of PE as a driver of organizationally relevant outcomes, including in-role and extra-role performance, turnover intentions, and PH indicators (Maynard et al., 2012, 2013; Seibert et al., 2011). Yet, these meta-analyses also reinforce the need for studies considering the PH-related outcomes of PE, and anchored in a true multilevel representation of PE. Although PE research has already a very rich multilevel tradition (Maynard et al., 2012, 2013; Seibert et al., 2011), this tradition remains limited, mainly, to two areas.

The first of those areas is team empowerment (Maynard et al., 2013), and the related concepts of team autonomy, efficacy, or potency (Gully, Incalterra, Joshi, & Beaubien, 2002; Hu & Liden, 2011; Van Mierlo, Rutte, Kompier, & Doorewaard, 2005). This research tradition is anchored in Kirkman and Rosen (1999) early theoretical developments who proposed that in a formal teamwork structure, team members' perceptions of their own PE needed to be differentiated from their perceptions of their team's empowerment (also see Bandura, 1997). To directly assess team empowerment, they (e.g., Kirkman, Rosen, Tesluk & Gibson, 2004) proposed to use items with a team referent (i.e., to create a climate measure based on the aforementioned distinction between context and climate). However, this research is difficult to generalize outside of formal teamwork structures, as shown by the fact that an attempt to transpose Kirkman et al.'s (2004) measure to the broader work unit context revealed very low estimates of inter-rater reliability (ICC2 = .29 to .52; D'Innocenzo et al., 2016). Yet, this research generated informative conclusions, showing that: (a) despite measurement *isomorphism*, individual and team PE measures shared differentiated associations with outcomes (Langfred, 2000), and interacted with one another (Chen et al., 2007), thus reinforcing the need for multilevel research; (b) inter-rater agreement (Arthur, Bell, & Edwards, 2007) and controlling for inter-rater agreement (i.e., seeking consensus among team members) (Kirkman, Tesluk, & Rosen, 2001) yielded more explanatory power, thus reinforcing the need to rely on latent aggregation methods. Yet, few studies have simultaneously considered two levels of analyses (Chen et al., 2007), or relied on a group-mean centering method to properly assess the climate effects of team empowerment (e.g., Kukenberger, Mathieu, & Ruddy, 2015), with most studies failing to even mention centering.

The second tradition adopted a multilevel approach to study PE antecedents and consequences (e.g., Chen et al., 2011; Liao, Toya, Lepak, & Hong, 2009; Seibert, Silver, & Randolph, 2004). Yet, this tradition has typically considered PE as an individual variable (L1), without seeking to understand its group-level (L2) effects, contextual or not. Still, assuming that PE could be predicted by L2 variables, and can influence L2 outcomes, implicitly recognizes the multilevel nature of PE. As noted by others, a group-level construct can only exert its influence at the group level (e.g., Zhang et al., 2009). Just like gender or achievement, PE is also likely to generate a specific empowerment context: personal feelings of competence, self-determination, meaning, and impact are not the same as exposure to a work environment where coworkers feel predominantly empowered or helpless. We can easily assume this work context to yield effects that are distinct from those of individual levels of PE. For example, exposure to highly empowered peers could, via social comparison processes, lead less empowered workers to question their work abilities. Conversely, via social learning processes, this same exposure could lead them to feel supported in their efforts to act in a volitional manner.

To our knowledge, only four studies have simultaneously analyzed PE at the individual and group level (Auh, Mengue, & Jung, 2014; Choi, 2007; Fong & Snape, 2015; Xu & Yang, 2018). Choi (2007)

found that PE had very similar associations with change-oriented organizational citizenship behaviors at L1 and L2. However, without an explicit mention of centering, it is hard to say whether these similar relations reflect a contextual effect (i.e., the L2 effect is already controlled for the L1 effect) or the opposite (i.e., once the L1 effect is subtracted from the L2 effect, nothing would be left). In a second study, Auh et al. (2014) reported slightly larger effects of PE on service-oriented citizenship behaviors at L2 than at L1. However, as it is based on group-mean centering, this result seems to suggest either a lack of contextual effect of PE, or a very small one. Fong and Snape (2015) reported group-mean centered results that differed across outcomes: (a) matching L1 and L2 effects of PE on job satisfaction (inconsistent with a contextual effect); (b) larger L2, relative to L1, effects of PE on in role behaviors (consistent with a positive contextual effect); (c) effects of PE limited to, or larger at, L1 in terms of organizational citizenship behaviors directed at individuals (consistent with a negative contextual effect); (d) large non-statistically significant effects of PE at L2 coupled with small non-statistically significant effects of PE at L1 (possibly consistent with a positive contextual effect).

Xu and Yang (2018) study is directly relevant to the present investigation as it relied on DL-MLSEM to study the associations between PE and emotional exhaustion at L1 and L2. Their results revealed negative effects of PE at both levels, although these effects were larger at L2. Unfortunately, given the authors' failure to mention centering, we have to assume that they relied on the group-mean centering inherent in DL-MLSEM. As such, their results are still consistent with a small contextual effect. Finally, an additional study considered the PE context at the group-level (L2) only (Wallace, Johnson, Mathe, & Paul, 2011), and found its effects to be statistically significant under conditions of high felt accountability. However, this pure L2 effect was not controlled for its L1 counterpart.

In sum, these studies provide very limited information on the likely role of the effects of the group-level PE context on employees' outcomes, particularly when PH is considered. Furthermore, although some studies suggested that PE might have a contextual effect, this effect generally seemed to be small and to vary as a function of the outcomes considered, in a way that was often obscured by improper centering strategies. Moreover, although most multilevel studies of team empowerment or PE report reliability estimates for L1 ratings (e.g., α or ω), and the L2 aggregation process (e.g., ICC2), we are aware of a single study that reported composite reliability estimates for L1 and L2 ratings simultaneously (ω_{L1} and ω_{L2}) (Langfred, 2005) and only two that controlled for these sources of errors via DL-MLSEM (D'Innocenzo et al., 2016; Xu & Yang, 2018). Likewise, although many studies tested the adequacy of preliminary measurement models, very few of them relied on multilevel measurement models for this verification (e.g., D'Innocenzo et al., 2016; Frazier & Fainshmidt, 2002; Langfred, 2005; Van Mierlo et al., 2007), and none tested *isomorphism* in factor loadings.

The Multilevel Structure of Psychological Health (PH) Assessments

The World Health Organization (2014) defines PH as a construct combining the presence of psychological well-being with the absence of psychological distress. This definition leaves open the question of whether psychological well-being and distress are best represented as the opposite endpoints of an overarching continuum or as distinct states able to fluctuate independently from one another. Evidence is rapidly accumulating for the first perspective based on studies realized among samples of youth (de Bruin & du Plessis, 2015; St Clair et al., 2017) and adults (Chen, Jing, Hayes, & Lee, 2013; Mu, Luo, Nickel, & Roberts, 2016) showing that PH is best represented as a bifactor construct with specific facets of psychological well-being and distress co-existing with a global factor mapping the overarching PH continuum. Although fewer studies have been conducted in the work area, their results also support this operationalization at the individual level (Laguna, Mielniczuk, & Razmus, 2019; Morin, Boudrias et al., 2016). To our knowledge, no research has yet considered the measurement structure of PH at the group level, or its possible contextual effects on various outcomes.

Matching recommendations made in PE research, calls have been made for a more systematic multilevel approach to the study of PH (Follmer & Jones, 2018; Martin, Karakina-Murray, Biron, & Sanderson, 2016). Yet, these calls have mainly focused on the operationalization of PH determinants or outcomes, and not on a true multilevel conception of PH. Yet, the mere recognition that PH determinants can both be located at the group or organization level, implies that PH is an inherently multilevel phenomenon. Indeed, just like PE, aggregated PH levels have a potential to generate contextual effects for employees exposed to either very high or very low collective levels of PH. Here also, social comparison processes, making one feel discouraged or encouraged by the PH levels displayed by others, or social learning or support processes, leading one to feel that this workplace is a

context that helps to nurture psychological well-being or distress, are likely to play a role.

Yet, emerging multilevel research on PH has generally focused on PH as an outcome variable (e.g., Marchand, Durand, Haines, & Harvey, 2015; Tucker, Sinclair, & Thomas, 2005; Tuckey, Bakker, & Dollard, 2012) and thus provides little information regarding the L1 and L2 outcomes of PH (for an exception, see Taris & Schreurs, 2009). Likewise, just as was the case in for PE, multilevel PH research is characterized by a lack of precision regarding centering decisions or centering decisions not aligned with the climate or contextual nature of the construct, a lack of information related to the reliability of multilevel measurement, few tests of multilevel measurement (Elovainio, Kivimäki, Steen, & Vahtera, 2004; Huhtala, Tolvanen, Mauno, & Feldt, 2015; Kiersch, & Byrne, 2015) not accompanied by verifications of *isomorphism*, and few attempts to control for measurement errors as part of the analyses (Elovainio et al., 2004; Huhtala et al., 2015; Kiersch, & Byrne, 2015).

In relation to the PE-PH association, a recent meta-analysis (Nielsen et al., 2017) of the key work-related determinants of PH, noted the key role of individual (L1) competence (self-efficacy) and collective (L2) self-determination (autonomy) and feelings of job control in the prediction of PH, thus suggesting effects occurring at more than one level of analysis (also see Elovainio, Kivimäki, Steen, & Kalliomäki-Levanto, 2000; Elovainio et al., 2004; Van Yperen & Snijders, 2000). These results thus add to those reported by Xu and Yang (2018) supporting multilevel effects of PE, reinforcing the idea that these effects may differ as a function of the outcome considered, as well as across PE dimensions.

References

Arthur, W., Bell, S.T., & Edwards, B.D. (2007). A longitudinal examination of the comparative criterion-related validity of additive and referent-shift consensus operationalization of team efficacy. *Organizational Research Methods*, *10*, 35-58.

Auh, S., Menguc, B., & Jung, Y.S. (2014). Unpacking the relationship between empowering leadership and service-oriented citizenship behaviors: A multilevel approach. *Journal of the Academy of Marketing Sciences*, *42*, 558-579.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Freeman.

Baynton, M.A., & Fournier, L. (2017). *The evolution of workplace mental health in Canada*. Toronto, Canada: The Great west Life Assurance Company.

Britt, T.W., & Dickinson, J.M. (2006). Morale during military operations: A positive psychology approach. In T.W. Britt, C.A. Castro, & A.B. Adler, (Eds.), *Military life: The psychology of serving in peace and combat* (pp. 157–184). Westport, CT: Praeger Security Int.

Brunetto, Y., Teo, S.T.T., Shaklock, K., & Farr-Wharton, R. (2012). Emotional intelligence, job satisfaction, well-being and engagement: Explaining organizational commitment and turnover intentions in policing. *Human Resource Management Journal*, *22*, 428-441.

Campbell, D.J., & Campbell, K.M. (2011). Impact of decision-making empowerment on attributions of leadership. *Military Psychology*, *23*, 154-179.

Canadian Standards Association Group, & Bureau de normalisation du Québec [CSA/BNQ] (2013). *Psychological health and safety in the workplace – Prevention, promotion, and guidance to staged implementation*. Ottawa, Canada: CSA Group. www.csagroup.org/documents/codes-and-standards/publications/CAN_CSA-Z1003-13_BNQ_9700-803_2013_EN.pdf

Chen, F.F., Jing, Y., Hayes, A., & Lee, J.M. (2013). Two concepts or two approaches? A bifactor analysis of psychological and subjective well-being. *Journal of Happiness Studies*, *14*, 1033-1068.

Chen, G., Kirkman, B.L., Kanfer, R., Allen, D., & Rosen, B. (2007). A multilevel study of leadership, empowerment, and performance in teams. *Journal of Applied Psychology*, *92*, 331-346.

Chen, G., Sharma, N., Edinger, S.K., Shapiro, D.L., & Farh, J.L. (2011). Motivating and demotivating forces in teams: Cross-level influences of empowering leadership and relationship conflict. *Journal of Applied Psychology*, *96*, 541-557.

Choi, J.N. (2007). Change-oriented organizational citizenship behaviors: Effects of work environment and intervening psychological processes. *Journal of Organizational Behavior*, *28*, 467-484.

Day, A., & Randell, K.D. (2014). Building a foundation for psychologically healthy workplaces and well-being. In A. Day, E.K. Kelloway, & J.J. Hurrell Jr. (Eds.), *Workplace well-being: How to build psychologically healthy workplaces* (pp. 3-26). Oxford, UK: Wiley-Blackwell.

deBruin, G.P., & du Plessis, G.A. (2015). Bifactor analysis of the mental health continuum-short form (MHC-SF). *Psychological Reports: Measures & Statistics*, *116*, 438-446.

Demerouti, E., Bakker, A.B., Vardakou, I., & Kantas, A. (2003). The convergent validity of two

burnout instruments: A multitrait-multimethod analysis. *European Journal of Psychological Assessment, 19*, 12–23.

D'Innocenzo, L., Luciano, M., Mathieu, J.E., Maynard, M.T., & Chen, G. (2016). Empowered to perform: A multilevel investigation of the influence of empowerment on performance in hospital units. *Academy of Management Journal*, *59*, 1290-1307.

Elovainio, M., Kivimäki, M., Steen, N., & Kalliomäki-Levanto, T. (2000). Organizational and individual factors affecting mental health and job satisfaction: A multilevel analysis of job control and personality. *Journal of Occupational Health Psychology*, *5*, 269-277.

Elovainio, M., Kivimäki, M., Steen, N., & Vahtera, J. (2004). Job decision latitude, organizational justice and health: Multilevel covariance structure analysis. *Social Science & Medicine, 58*, 1659-1669.

Follmer, K.B., & Jones, K.S. (2018). Mental illness in the workplace: An interdisciplinary review and organisational research agenda. *Journal of Management*, *44*, 325-351.

Fong, K., & Snape, E. (2015). Empowering leadership, psychological empowerment and employee outcomes: Testing a multi-level mediating model. *British Journal of Management*, *26*, 126-138.

Frazier, M.L., & Fainshmidt, S. (2012). Voice climate, work outcomes, and the mediating role of psychological empowerment. *Group & Organization Management*, *37*, 691-715

Griffeth, R.W., Hom, P., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of Management*, *26*, 463-488.

Gully, S.M., Incalcaterra, K.A., Joshi, A., & Beaubien, J.M. (2002). A meta-analysis of team efficacy, potency, and performance: Interdependence and level of analysis as moderators of observed relationships. *Journal of Applied Psychology*, *87*, 819-832.

Hancock, J.I., Allen, D.G., Bosco, F.A., McDaniel, K.R., & Pierce, C.A. (2013). Meta-analytic review of employee turnover as a predictor of firm performance. *Journal of Management*, *39*, 573-603.

Harter, J.K., Schmidt, F.L., & Keyes, C.L.M. (2003). Well-being in the workplace and its relationship to business outcomes: A review of the Gallup studies. In C.L.M. Keyes & J. Haidt (Eds.), *Flourishing: Positive psychology and the life well-lived* (pp. 205-224). Washington, DC: American Psychological Association.

Heavey, A.L., Holwerda, J.A., & Hausknecht, J.P (2013). Causes and consequences of collective turnover: A meta-analytic review. *Journal of Applied Psychology*, *98*, 412-453.

Hom, P.W., Lee, T.W., Shaw, J.D., & Hausknecht, J.P. (2017). One hundred years of employee turnover theory and research. *Journal of Applied Psychology*, *102*, 530-545.

Hu, J., & Liden, R. C. (2011). Antecedents of team potency and team effectiveness: An examination of goal and process clarity and servant leadership. *Journal of Applied psychology*, *96*, 851-862.

Huhtala, M., Tolvanen, A., Mauno, S., & Feldt, T. (2015). The associations between ethical organizational culture, burnout, and engagement: A multilevel study. *Journal of Business & Psychology*, *30*, 399-414.

Ivey, G.W., Blanc, J.R.S., Michaud, K., & Dobreva-Martinova, T. (2018). A measure and model of psychological health and safety in the workplace that reflects Canada's national standard. *Canadian Journal of Administrative Sciences*, *35*, 509-522.

Kessler, R.C., Andrews, G., Colpe, L.J., Hiripi, E., Mroczek, D.K., Normand, S.-L.T., Walters, E.E., Zaslavsky, A.M. (2002). Short screening scales to monitor population prevalence and trends in non-specific psychological distress. *Psychological Medicine*, *32*, 959–976.

Kiersch, K.E., & Byrne, Z.S. (2015). Is being authentic being fair? Multilevel examination of authentic leadership, justice, and outcomes. *Journal of Leadership & Organizational Studies*, *22*, 292-303.

Kirkman, B.L., & Rosen, D. (1999). Beyond self-management: Antecedents and consequences of team empowerment. *Academy of Management Journal*, *42*, 58-74.

Kirkman, B.L., Rosen, D., Tesluk, P.E., & Benson, C.B. (2004). The impact of team empowerment on virtual team performance: The moderating role of face-to-face interactions. *Academy of Management Journal*, *47*, 175-192.

Kirkman, B.L., Tesluk, P.E., & Rosen, D. (2001). Assessing the incremental validity of team consensus ratings over aggregation of individual-level data in predicting team effectiveness. *Personnel Psychology*, *54*, 645-667.

Kotze, E., Menon, S.T., & Vos, B. (2007). Psychological empowerment in the South African military. *South African Journal of Industrial Psychology*, *33*, 1-6.

Kukenberger, M.R., Mathieu, J.E., & Ruddy, T. (2015). A cross-level test of empowerment and process influences on members' informal learning and team commitment. *Journal of Management, 41,* 987-1016.

Laguna, M., Mielniczuk, E., & Razmus, W. (2019). Test of the bifactor model of job-related affective well-being. *Europe's Journal of Psychology, 15,* 342-357.

Langfred, C.W. (2000). The paradox of self-management: Individual and group autonomy in work groups. *Journal of Organizational Behavior, 21,* 563-585.

Langfred, C.W. (2005). Autonomy and performance in teams: The multilevel moderating effect of task interdependence. *Journal of Management, 31,* 513-529.

Liao, H., Toya, K., Lepak, D.P., & Hong, Y. (2009). Do they see eye to eye? Management and employee perspectives of high-performance work systems and influence processes on service quality. *Journal of Applied Psychology, 94,* 371-391.

Lytell, M.C., & Drasgow, F. (2009)."Timely" methods: Examining turnover rates in the U.S. military. *Military Psychology, 21,* 334-350

Marchand, A., Durand, P., Haines, V., & Harvey, S. (2015). The multilevel determinants of workers' mental health. *Social Psychiatry & Psychiatric Epidemiology, 50,* 445-459.

Martin, A., Karanika-Murray, M., Biron, C., & Sanderson, K. (2016). The psychosocial work environment, employee mental health and organizational interventions. *Stress & Health, 32,* 201-215.

Maynard, M.T., Gilson, L.L., & Mathieu, J.E. (2012). Empowerment – Fad or Fab? A multilevel review of the past two decades of research. *Journal of Management, 38,* 1231-1281,

Maynard, M.T., Mathieu, J.E., Gilson, L.L., O'Boyle, E.H., & Cigularov, K.P. (2013). Drivers and outcomes of team psychological empowerment: A meta-analytic review and model test. *Organizational Psychology Review, 3,* 101-137

Mor Barak, M.E., Nissly, J.A., & Levin, A. (2001). Antecedents to retention and turnover among child welfare, social work, and other human service employees. *Social Service Review, 75,* 625-661.

Morin, A.J.S., Boudrias, J.-S., Marsh, H.W., Madore, I., & Desrumaux, P. (2016). Further reflections on disentangling shape and level effects in person-centered analyses: An illustration exploring the dimensionality of psychological health. *Structural Equation Modeling, 23,* 438-454.

Morin, A.J.S., Meyer, J.P., Bélanger, É., Boudrias, J.-S., Gagné, M., & Parker, P.D. (2016). Longitudinal associations between employees' perceptions of the quality of the change management process, affective commitment to change and psychological empowerment. *Human Relations, 69,* 839-867.

Mu, W., Luo, J., Nickel, L., & Roberts, B.W. (2016). Generality or specificity? Examining the relation between personality traits and mental health outcomes using a bivariate bi-factor latent change model. *European Journal of Personality, 30,* 467-483.

Nielsen, K., Nielsen, M.B., Ogbonnaya, C., Känsälä, M., Saari, E., & Isaksson, K. (2017). Workplace resources to improve both employee well-being and performance: A systematic review and meta-analysis. *Work & Stress, 31,* 101-120.

Park, T.-Y., & Shaw, J.D. (2013). Turnover rates and organizational performance: A meta-analysis. *Journal of Applied Psychology, 98,* 268-309.

Rasmussen, M.F. (2012). *A framework of organizational empowerment for strategic military leaders (strategy research project #ADA561792)*. Carlisle, PA: Defense Technical Information Center, United States Army War College.

Raymer, S.R., Lindsay, D., & Watola, D.J. (2017). A systematic approach for building alignment: Establishing responsibilities and opportunities for recognizing role development and diminishing stress. In A. MacIntyre, D., Lagacé-Roy, & D. Lindsay (Eds.). *Global views on military stress and resilience* (pp. 151-170). Kingston, Ontario: Canadian Defence Academy Press.

Rich, B.L., Lepine, J.A., & Crawford, E.R. (2010). Job engagement: Antecedents and effects on job performance. *Academy of Management Journal, 53,* 617–635.

Rubenstein, A.L., Eberly, M.B., Lee, T.W., & Mitchell, T.R. (2018). Surveying the forest: A meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover. *Personnel Psychology, 71,* 23-65.

Seibert, S.E., Silver, S.R., & Randolph, W.A. (2004). Taking empowerment to the next level: A multiple-level model of empowerment. *Academy of Management Journal, 47,* 332-349.

Seibert, S.E., Wang, G., & Courtright, S.H. (2011). Antecedents and consequences of psychological and team empowerment in organizations. *Journal of Applied Psychology, 96,* 981-1003.

Spreitzer, G.M. (1995). Psychological empowerment in the workplace: Dimensions, measurement, and validation. *Academy of Management Journal, 38,* 1442-1465.

Spreitzer, G.M. (2008). Taking Stock: A review of more than twenty years of research on empowerment at work. In C. Cooper & J. Barling (Eds.), *Handbook of organizational behavior* (pp. 54-73). Thousand Oaks, CA: Sage.

St Clair, M.C., Neufeld, S., Jones, P.B., Fonagy, P., Bullmore, E.T., Dolan, R.J., Moutoussis, M., Toseeb, U., & Goodyer, I.M. (2017). Characterizing the latent structure and organisation of self-reported thoughts, feelings and behaviours in adolescents and young adults. *Plos One, 12,* e0175381.

Taris, T.W., & Schreurs, P.J.G. (2009). Well-being and organizational performance: An organizational-level test of the happy-productive worker hypothesis. *Work & Stress, 23,* 120-136.

Thomas, K.W., & Velthouse, B.A. (1990). Cognitive elements of empowerment: An "interpretive" model of intrinsic task motivation. *Academy of Management Review, 15,* 666-681.

Tucker, J.S., Sinclair, R.R., & Thomas, J.L. (2005). The multilevel effects of occupational stressors on soldier's well-being, organizational attachment, and readiness. *Journal of Occupational Health Psychology, 10,* 276-299.

Tuckey, M.R., Bakker, A.B., & Dollard, M.F. (2012). Empowering leaders optimize working conditions for engagement: A multilevel study. *Journal of Organizational Health Psychology, 17,* 15-27.

Van Yperen, N.W., & Snijders, T.A.B. (2000). A multilevel analysis of the demands-control model: Is stress at work determined by factors at the group level or the individual level? *Journal of Occupational Health Psychology, 5,* 182-190.

Van Mierlo, H., Rutte, C.G., Kompier, M.A.J., & Doorewaard, H.A.C.M. (2005). Self-managing teamwork and psychological well-being: A review of a multilevel research domain. *Group & Organization Management, 30,* 211-235.

Van Mierlo, H., Rutte, C.G., Vermunt, J.K., Kompier, M.A.J., & Doorewaard, J.A.C.M. (2007). A multi-level mediation model of the relationships between team autonomy, individual task design and psychological well-being. *Journal of Occupational & Organizational Psychology, 80,* 647-664.

Wallace, J.C., Johnson, P.D., Mathe, K., & Paul, J. (2011). Structural and psychological empowerment climates, performance, and the moderating role of shared felt accountability: A managerial perspective. *Journal of Applied Psychology, 96,* 840-850.

World Health Organization (2014). *Mental health*. www.who.int/features/factfiles/mental_health/en/.

Wu, W., Rafiq, M., & Chin, T. (2017). Employee well-being and turnover intention. *Career Development International, 22,* 797-815.

Xu, Z., & Yang, F. (2018). The cross-level effect of authentic leadership on teacher emotional exhaustion. *Journal of Pacific Rim Psychology, 12,* e35.

Yanchus, N.J., Periard, D., & Osatuke, K. (2017). Further examination of predictors of turnover intention among mental health professionals. *Journal of Psychiatric & Mental Health Nursing, 24,* 41-56.

Zhang, Z., Zyphur, M.J., & Preacher, K.J. (2009). Testing multilevel mediation using hierarchical linear models problems and solutions. *Organizational Research Methods,* 12, 695-719.

Section 3

Theoretical Discussion of Multilevel Relations among PE, PH, and Turnover Intentions

The results from the analyses reported in the main manuscript have implications for research in PE, PH, and turnover intentions in relation to multilevel measurement and prediction.

PE Measurement. Research conducted on the PE construct has generally established, in a very convincing manner, that PE is best represented as a higher-order construct formed of four dimensions reflecting feelings of competence, self-determination, impact, and meaning (Seibert et al., 2011; Maynard et al., 2012, 2013). Yet, despite the generic recognition of *isomorphism* between measures of PE and team empowerment (Seibert et al., 2011), true multilevel tests of the extent to which this structure generalizes across levels for measures of individual PE analyzed at the item level are rare (Frazier & Fainshmidt, 2002). Importantly, higher-order factor models are limited by the inclusion of a strict proportionality constraint in which the ratio of higher-order to first-order variance explained at the item level is forced to be constant across the items forming a scale (Gignac, 2016), reinforcing the importance of contrasting more parsimonious higher-order factor models with flexible bifactor models (Gignac, 2016; Reise, 2012). In this regard, the present study was informative in demonstrating that PE seemed to represent one of the few *true* higher-order constructs, as well as a construct whose structure fully generalized across levels. Furthermore, whereas measures of team empowerment makes them hard to generalize outside of formal teamwork structure (as shown by the weak ICC2 obtained by D'Innocenzo et al., 2016), our results showed individual PE measures could be reliably aggregated at the group level (ω_{L2} = .848 to .998; ICC2 = .822 to .909).

PH Measurement. Contrasting with PE, the measurement structure of PH matched a bifactor representation where each item directly reflected a global PH construct characterized by negative loadings from the distress items and positive loadings from the well-being items, as well as co-existing specific PH factors reflecting specificity left unexplained by the global factor. These results not only provided replication evidence for previous results obtained in relation to the measurement of PH at work (Laguna et al., 2019; Morin, Boudrias et al., 2016), but did so using a collection of different PH measures. The present study was also the first to document the measurement *isomorphism* of PH at both the individual and work unit level. Although some of the S-factors estimated as part of this model were found to retain only a limited amount of specificity at L1 (emotional exhaustion and disengagement), L2 (disengagement), or in terms of inter-rater agreement (cognitive engagement: which is a highly individual and internalized construct) once the PH G-factor was taken into account, it was impressive to note that the majority of these S-factors proved to be highly reliable across indicators (morale, physical engagement, emotional engagement, and psychological distress). In addition, the global PH factor itself proved to be a reliable indicator of employees' individual PH levels (ω_{L1} = .976) as well as work units aggregated PH levels (ω_{L2} = .995; ICC2 = .892).

Associations between PE, PH, and Turnover Intentions. Given our objective to illustrate DL-MLSEM procedures, we selected constructs for which the associations were very well-documented at the individual level (PE→PH/turnover intentions: Maynard et al., 2012; Nielsen et al., 2017; Seibert et al., 2011; PH→turnover intentions: Mor Barak et al., 2001; Rubenstein et al., 2018) but undocumented at the work unit level. This verification sought to answer repeated calls for increases in multilevel research focusing on these constructs and their associations (e.g., Follmer & Jones, 2018; Hom et al., 2017; Martin et al., 2016; Rubenstein et al., 2018), and to verify their possible contextual nature.

Supporting the adequacy of our measures and attesting to the generalizability of our findings, the results supported our a priori predictive model at the individual level, illustrating the role of PH as a partial mediator of the relation between PE and turnover intentions, with residual direct effects of PE on turnover intentions also matching our expectations. Furthermore, most of these associations were also apparent at the work unit level, albeit slightly diluted by the aggregation of individual effects at the work unit level. More precisely, the strength of these associations became weaker at the work unit level, where the residual direct effect of PE on turnover intention was no longer present, thus showing evidence of full mediation at the group level. These results generally support the value of relying on this test battery for purposes of assessments conducted at the work unit level, as group-level effects of PE on PH, and of PH on turnover intentions remained statistically significant.

However, we found no evidence that either PE or PH yielded any additional contextual effects

once aggregated individual effects were considered. This lack of contextual effect does not imply that these variables have no L2 effect, simply that they do not lead to the formation of a specific group context that, in and of itself, impacts group outcomes beyond the effect already created by the individual levels of PE and PH of group members. Such as contextual effect would have meant that the simple exposure to highly or weakly empowered or healthy co-workers could result in additional effects beyond the effects of employees' individual levels of PE or PH, invoking possible social comparisons or learning processes (e.g., Marsh et al., 2012; Morin et al., 2014). This did not happen in this study, at least as far as turnover intentions were considered. Although disappointing, this observation still adds to our understanding of both constructs as they are expressed at the group level, and help to shed clarity on previous results which, due to variations in centering strategies, could be alternatively interpreted as supporting, or not, the presence of contextual effects (e.g., Auh et al., 2014; Choi, 2007; Fong & Snape, 2015; Xu & Yang, 2018). Clearly, additional research will be needed to empirically verify whether these conclusions generalize to other outcomes.

Another important contribution of this study was the observation that most associations, at both L1 and L2, were happening at the level of the global PE and PH construct. Yet, additional associations were also observed at the subscale level. Thus, the self-determination and competence PE dimensions played a role, beyond that of global PE, at the individual level in the prediction of morale, disengagement, psychological distress and emotional engagement. One of these effects was strong enough to be also observable at the aggregate group level (the effect of self-determination on disengagement). In contrast, at the group level, it was the meaning facet of PE that played a specific role in the prediction of emotional and cognitive engagement. Although these raw L2 effects did not translate into contextual effects, this could possibly be related to the lower level of inter-rater reliability associated with these job engagement dimensions.

Importantly, these additional results are consistent with previous studies which have also demonstrated that different PE dimensions tended to exert distinct effects at the group relative to individual level (Elovainio et al., 2000, 2004; Nielsen et al., 2017; Van Yperen & Snijders, 2000). Yet, although these previous studies also found marked effects of competence at the individual level, they found self-determination to play a greater role at the group level, a result which is inconsistent with the present results. It is important to keep in mind that these previous studies did not focus on the global, four-dimensional, PE construct but on isolated measures of competence (or self-efficacy) and self-determination (or autonomy) often considered in isolation. It is thus hard to transpose their results to the present context, particularly given their failure to also consider employees global PE levels as a key driver of outcome associations. Clearly, this is also an area where additional research is needed.

Limitations and Directions for Future Research. Despite the many strengths of the present study, it still presents a few noteworthy limitations that should be kept in mind when considering its substantive implications. First, the present study remains cross sectional in nature, and thus unable to clearly disaggregate the directionality of the observed associations among PE, PH, and turnover intentions. Although the directionality of these associations implied in the model that we tested was theoretically grounded in a well-researched area, this is not sufficient to overshadow the advantages of longitudinal research which is able to provide tests of directionality, but also direct estimates of stability, changes, and trajectories in the constructs of interest (Grimm, Ram, & Estabrook, 2016). Importantly, some recent calls have been made to increase researchers' attention to the implication of time-sensitive within-person fluctuations in PH levels (Bakker, 2015; Ilies, Aw, & Pluut, 2015), which would be a very interesting area of investigation to combine with multilevel considerations, allowing one to study group impacts on individual trajectories. Second, despite its reliance on a very large multilevel sample and the demonstration that the estimated measurement models generalized to males and females, military and civilian employees, and French or English-speaking respondents, the generalizability of the present results beyond the Canadian defence context needs replication.

Third, although the current results showed a lack of contextual effects associated with PE and PH, it is important to keep in mind that these results are limited to the effects of PE and PH on the outcomes considered here (i.e., PH and turnover intentions). Importantly, turnover intentions only presented a very limited level of variability at the between-group level (ICC1= .024) which might have made it harder to detect true contextual effects. The fact that raw L2 effects could be identified reinforce the idea that this smaller amount of group-level variance was not enough to preclude the detection of L2 effects. Yet, it also reinforces the need to replicate the present study while considering

broader range of outcome variables presenting more between-group variability, or reflecting pure L2 constructs (e.g., official ratings of unit-level productivity). A last limitation stems from our sole reliance on self-reported measures. Fortunately, as shown in Siemsen, Roth, and Oliveira's (2010) mathematical demonstration, shared method variance is unlikely to play a role in the context of multivariate analyses, in addition to being naturally controlled for in DL-MLSEM (where group-mean centering re-expresses all scores as deviation from the group mean on a specific construct). Yet, it would have been informative to combine these measures with more objective measures of empowerment (such as supervisors' ratings of employees' empowered behaviors; Morin et al., 2011) or health (such as objective measures of absenteeism or sick leaves).

References

Auh, S., Menguc, B., & Jung, Y.S. (2014). Unpacking the relationship between empowering leadership and service-oriented citizenship behaviors: A multilevel approach. *Journal of the Academy of Marketing Sciences, 42*, 558-579.

Bakker, A.B. (2015). Towards a multilevel approach of employee well-being. *European Journal of Work & Organizational Psychology, 24*, 839-843.

Choi, J.N. (2007). Change-oriented organizational citizenship behaviors: Effects of work environment and intervening psychological processes. *Journal of Organizational Behavior, 28*, 467-484.

D'Innocenzo, L., Luciano, M., Mathieu, J.E., Maynard, M.T., & Chen, G. (2016). Empowered to perform: A multilevel investigation of the influence of empowerment on performance in hospital units. *Academy of Management Journal*, *59*, 1290-1307.

Elovainio, M., Kivimäki, M., Steen, N., & Kalliomäki-Levanto, T. (2000). Organizational and individual factors affecting mental health and job satisfaction: A multilevel analysis of job control and personality. *Journal of Occupational Health Psychology, 5*, 269-277.

Elovainio, M., Kivimäki, M., Steen, N., & Vahtera, J. (2004). Job decision latitude, organizational justice and health: Multilevel covariance structure analysis. *Social Science & Medicine, 58*, 1659-1669.

Follmer, K.B., & Jones, K.S. (2018). Mental illness in the workplace: An interdisciplinary review and organisational research agenda. *Journal of Management*, *44*, 325-351.

Fong, K., & Snape, E. (2015). Empowering leadership, psychological empowerment and employee outcomes: Testing a multi-level mediating model. *British Journal of Management, 26*, 126-138.

Frazier, M.L., & Fainshmidt, S. (2012). Voice climate, work outcomes, and the mediating role of psychological empowerment. *Group & Organization Management*, *37*, 691-715

Gignac, G.E. (2016). The higher-order model imposes a proportionality constraint: That is why the bifactor model tends to fit better. *Intelligence, 55*, 57–68.

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches*. New York, NY: Guilford.

Hom, P.W., Lee, T.W., Shaw, J.D., & Hausknecht, J.P. (2017). One hundred years of employee turnover theory and research. *Journal of Applied Psychology, 102*, 530-545.

Ilies, R., Aw, S.Y, & Pluut, H. (2015). Intraindividual models of employee well-being. *European Journal of Work & Organizational Psychology, 24*, 827-838

Laguna, M., Mielniczuk, E., & Razmus, W. (2019). Test of the bifactor model of job-related affective well-being. *Europe's Journal of Psychology*, *15*, 342-357.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S. & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*, 106-124.

Martin, A., Karanika-Murray, M., Biron, C., & Sanderson, K. (2016). The psychosocial work environment, employee mental health and organizational interventions. *Stress & Health*, *32*, 201-215.

Maynard, M.T., Gilson, L.L., & Mathieu, J.E. (2012). Empowerment – Fad or Fab? A multilevel review of the past two decades of research. *Journal of Management, 38*, 1231-1281,

Maynard, M.T., Mathieu, J.E., Gilson, L.L., O'Boyle, E.H., & Cigularov, K.P. (2013). Drivers and outcomes of team psychological empowerment: A meta-analytic review and model test. *Organizational Psychology Review, 3*, 101-137

Mor Barak, M.E., Nissly, J.A., & Levin, A. (2001). Antecedents to retention and turnover among child welfare, social work, and other human service employees. *Social Service Review*, *75*, 625-661.

Morin, A.J.S., Boudrias, J.-S., Marsh, H.W., Madore, I., & Desrumaux, P. (2016). Further reflections on disentangling shape and level effects in person-centered analyses: An illustration exploring the

dimensionality of psychological health. *Structural Equation Modeling*, *23*, 438-454.

Morin, A.J.S., Marsh, H.W., Nagengast, B., & Scalas, L.F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *The Journal of Experimental Education*, *82*, 143-167.

Morin, A.J.S., Vandenberghe, C., Boudrias, J.-S., Madore, I., Morizot, J., & Tremblay, M. (2011). affective commitment and citizenship behaviors across multiple foci. *Journal of Managerial Psychology*, *26*, 716-738

Nielsen, K., Nielsen, M.B., Ogbonnaya, C., Känsälä, M., Saari, E., & Isaksson, K. (2017). Workplace resources to improve both employee well-being and performance: A systematic review and meta-analysis. *Work & Stress*, *31*, 101-120.

Reise, S.P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696.

Rubenstein, A.L., Eberly, M.B., Lee, T.W., & Mitchell, T.R. (2018). Surveying the forest: A meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover. *Personnel Psychology*, *71*, 23-65.

Seibert, S.E., Wang, G., & Courtright, S.H. (2011). Antecedents and consequences of psychological and team empowerment in organizations. *Journal of Applied Psychology*, *96*, 981-1003.

Siemsen, E., Roth, A., & Oliveira, P. (2010). Common method bias in regression models with linear, quadratic, and interaction effects. *Organizational Research Methods*, *13*, 456-476.

Van Yperen, N.W., & Snijders, T.A.B. (2000). A multilevel analysis of the demands-control model: Is stress at work determined by factors at the group level or the individual level? *Journal of Occupational Health Psychology*, *5*, 182-190.

Xu, Z., & Yang, F. (2018). The cross-level effect of authentic leadership on teacher emotional exhaustion. *Journal of Pacific Rim Psychology*, *12*, e35.

Section 4

Additional Results from the Measurement Models Discussed in the Main Manuscript

Table S1

Standardized Factor Loadings (λ) and Uniquenesses (δ) for the Empowerment and Turnover Intentions Measurement Models

	Single Level		Multilevel: Level 1		Multilevel: Level 2	
Items	λ	δ	λ	δ	λ	δ
Turnover Intentions						
Item 1	.678	.541	.674	.546	.661	.563
Item 2	.805	.352	.810	.345	.683	.534
Item 3	.863	.255	.862	.256	.927	*.140*
Self-Determination (First-Order)						
Item 1	.683	.534	.670	.551	.950	*.098*
Item 2	.370	.863	.355	.874	.447	.800
Item 3	.614	.623	.599	.641	.929	*.136*
Item 4	.647	.582	.636	.596	.837	.299
Item 5	.801	.359	.794	.370	.932	*.132*
Item 6	.626	.608	.609	.629	.873	.238
Impact (First-Order)						
Item 1	.780	.392	.775	.400	.963	*.072*
Item 2	.937	.121	.936	.124	.983	*.034*
Item 3	.946	.106	.943	.110	.835	.302
Meaning (First Order)						
Item 1	.933	.129	.931	.133	.995	*.010*
Item 2	.936	.124	.934	.128	.999	*.002*
Item 3	.971	.057	.970	.059	.997	*.006*
Competence (First-Order)						
Item 1	.793	.371	.790	.377	.850	.277
Item 2	.861	.258	.857	.265	.989	*.022*
Item 3	.842	.291	.839	.296	.971	*.058*
Item 4	.784	.385	.782	.389	.851	.277
Empowerment (Higher-Order)						
Autonomy	.882	.222	.882	.223	.807	.349
Impact	.739	.454	.744	.447	.816	*.334*
Meaning	.664	.559	.653	.574	.917	*.159*
Competence	.363	.868	.349	.878	.470	.779

Note. λ: factor loading; δ: item uniqueness; parameter estimates that are non-statistically significant ($p > .05$) are marked in italics.

Table S2

Standardized Factor Loadings (λ) and Uniquenesses (δ) for the Psychological Health Bifactor Measurement Models

	Single Level			Multilevel: Level 1			Multilevel: Level 2		
Items	S-λ	G- λ	δ	S-λ	G- λ	δ	S-λ	G- λ	δ
Morale									
Item 1	.519	.670	.282	.529	.722	.288	.443	.828	.095
Item 2	.355	.726	.346	.351	.658	.355	.264	.842	.245
Item 3	.405	.693	.355	.404	.687	.365	.342	.888	.094
Item 4	.581	.667	.218	.585	.657	.227	.504	.863	*.000*
Item 5	.592	.696	.165	.595	.688	.172	.491	.864	*.012*
Item 6	.602	.647	.219	.604	.640	.226	.502	.811	*.091*
Physical Engagement									
Item 1	.700	.350	.387	.697	.352	.391	.569	.681	*.212*
Item 2	.787	.370	.244	.788	.366	.245	.672	.739	*.002*
Item 3	.828	.254	.250	.824	.269	.249	.602	.466	.420
Item 4	.730	.337	.353	.731	.333	.355	.674	.728	*.016*
Item 5	.784	.173	.356	.782	.183	.356	.493	.272	.683
Item 6	.735	.345	.341	.737	.340	.341	.660	.721	*.044*
Emotional Engagement									
Item 1	.522	.706	.229	.524	.701	.234	.327	.931	*.026*
Item 2	.372	.765	.276	.372	.761	.282	.215	.935	*.080*
Item 3	.570	.626	.284	.573	.619	.289	.393	.902	*.032*
Item 4	.569	.617	.295	.571	.611	.301	.400	.911	*.009*
Item 5	.480	.765	.184	.484	.759	.190	.286	.951	*.015*
Item 6	.548	.719	.183	.551	.712	.189	.305	.838	.206
Cognitive Engagement									
Item 1	.575	.582	.330	.578	.574	.336	.308	.951	*.000*
Item 2	.769	.492	.167	.772	.485	.169	.454	.888	*.006*
Item 3	.807	.461	.136	.810	.454	.138	.401	.700	.349
Item 4	.643	.336	.473	.643	.337	.473	.420	.685	*.354*
Item 5	.746	.492	.200	.749	.486	.203	.443	.895	*.003*
Item 6	.802	.442	.162	.803	.437	.163	.507	.859	*.006*
Exhaustion									
Item 1	.225	-.598	.592	.229	-.588	.602	*.137*	-.875	.216
Item 2	.356	-.744	.320	.357	-.733	.334	*.188*	-.956	*.050*
Item 3	.252	-.760	.359	.252	-.754	.368	*.126*	-.937	.106
Item 4	.452	-.709	.293	.469	-.699	.291	*.240*	-.886	.158
Disengagement									
Item 1	.162	-.777	.371	.160	-.767	.387	.190	-.916	.126
Item 2	.363	-.604	.503	.358	-.592	.521	.473	-.791	*.151*
Item 3	.460	-.607	.421	.455	-.594	.440	.602	-.794	*.008*
Item 4	.199	-.671	.511	.190	-.663	.524	.226	-.798	.313
Psychological Distress									
Item 1	.192	-.705	.466	.193	-.695	.480	.140	-.952	*.075*
Item 2	.411	-.519	.562	.408	-.513	.570	.309	-.738	.359
Item 3	.557	-.480	.459	.558	-.471	.467	.516	-.827	*.050*
Item 4	.558	-.636	.284	.561	-.627	.292	.416	-.882	*.049*
Item 5	.372	-.565	.543	.374	-.554	.553	.286	-.804	.271
Item 6	.423	-.494	.576	.425	-.483	.586	.359	-.775	.271
Item 7	.580	-.645	.248	.584	-.636	.255	.430	-.889	*.024*
Item 8	.410	-.672	.380	.413	-.662	.391	.304	-.925	*.052*
Item 9	.672	-.555	.240	.676	-.547	.245	.529	-.812	*.060*
Item 10	.577	-.547	.368	.582	-.536	.374	.461	-.805	.140

Note. G: Global factor from a bifactor measurement model; S: Specific factor from a bifactor measurement model; λ: factor loading; δ: item uniqueness; parameter estimates that are non-statistically significant ($p > .05$) are marked in italics.

Table S3

Latent Variable Correlations Estimated as Part of the Overall Multilevel Measurement Model (Level 1 Correlations Below the Diagonal, and Level 2 Correlations Above the Diagonal)

		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.
1.	Self-Determination (First-Order)		0	0	0	0	-.529	.917	.210	-.224	-.290	-.358	-.067	-.372	*.000*
2.	Impact (First-Order)	0		0	0	0	-.446	.560	.477	.042	-.127	-.225	*.003*	.028	.307
3.	Meaning (First Order)	0	0		0	0	-.525	.811	.370	-.073	-.064	-.197	*.017*	-.238	.110
4.	Competence (First-Order)	0	0	0		0	-.346	.568	.288	-.052	-.031	-.162	-.089	-.061	.030
5.	Empowerment (Higher-Order)	0	0	0	0		-.905	.992	*.035*	*-.006*	*.104*	*-.237*	*.188*	*-.098*	*.087*
6.	Turnover Intentions	-.527	-.336	-.475	-.173	-.685		-.914	*-.337*	*.257*	*-.142*	*-.173*	*.070*	*.165*	*.042*
7.	Global Psychological Health (G-Factor)	.677	.459	.604	.277	.820	-.738		0	0	0	0	0	0	0
8.	Morale (S-factor)	.244	.216	.252	.179	.150	*-.080*	0		0	0	0	0	0	0
9.	Physical Engagement (S-factor)	.037	.074	.166	.154	-.070	.081	0	0		0	0	0	0	0
10.	Emotional Engagement (S-factor)	.138	.134	.335	.130	.234	-.150	0	0	0		0	0	0	0
11.	Cognitive Engagement (S-factor)	.050	.087	.197	.114	*-.014*	*.030*	0	0	0	0		0	0	0
12.	Exhaustion (S-factor)	.056	.114	.119	.074	.190	-.119	0	0	0	0	0		0	0
13.	Disengagement (S-factor)	-.177	-.154	-.156	.002	*-.143*	*.171*	0	0	0	0	0	0		0
14.	Psychological Distress (S-factor)	-.058	-.049	-0.024	-.079	*-.009*	*.032*	0	0	0	0	0	0	0	
	ω	.797	.920	.963	.892	.769	.827	.977	.855	.915	.866	.928	.514	.437	.846
	ω_{L1}	.786	.917	.962	.889	.765	.828	.976	.852	.915	.864	.928	.517	.419	.844
	ω_{L2}	.935	.950	.998	.955	.848	.807	.995	.923	.907	.910	.899	.474	.788	.912
	ICC1	.077	.044	.051	.037	.064	.024	.065	.029	.012	.014	.006	.019	.077	.027
	ICC2	.909	.847	.865	.822	.892	.750	.892	.782	.600	.638	.420	.702	.909	.769

Note. G: Global factor from a bifactor measurement model; S: Specific factor from a bifactor measurement model; ω: omega coefficient of model-based composite reliability estimated from a single level model; ω_{L1}: omega coefficient of model-based composite reliability estimated at level 1 (individual) from a multilevel model; ω_{L2}: omega coefficient of model-based composite reliability estimated at level 2 (unit) from a multilevel model; ICC1: Intraclass correlation; ICC2: Reliability of unit-level aggregations; parameter estimates that are non-statistically significant ($p > .05$) are marked in italics; The number 0, with no decimal, mark correlations not estimated as part of the model.

Table S4

Goodness-of-Fit Statistics of the Single Level Measurement Invariance Models

Description	χ^2 (*df*)	CFI	TLI	RMSEA	CM	$\Delta\chi^2$ (*df*)	ΔCFI	ΔTLI	ΔRMSEA
English versus French Version									
F1. Configural invariance (First-Order)	17167.766 (3344)*	.928	.921	.038	--	--	--	--	--
F2. Weak invariance (First-Order)	17597.302 (3434)*	.926	.921	.038	F1	435.735 (90)*	-.002	.000	.000
F3. Strong invariance (First-Order)	18979.470 (3482)*	.919	.915	.040	F2	1318.693 (48)*	-.007	-.006	+.002
F4. Strict invariance(First-Order)	19504.044 (3543)*	.917	.914	.040	F3	427.863 (61)*	-.002	-.001	.000
F5. Correlated Uniq. (First-Order)	19585.330 (3548)*	.916	.914	.040	F4	104.780 (5)*	-.001	.000	.000
F6. Latent Var.-Covar. (First-Order)	19624.825 (3606)*	.916	.915	.040	F5	120.277 (58)*	.000	+.001	.000
F7. Latent Means (First-Order)	20033.481 (3619)*	.914	.913	.040	F6	285.778 (13)*	-.002	-.002	.000
S1. Configural invariance (Second-Order)	20877.202 (3631)*	.910	.909	.041	--	--	--	--	--
S2. Weak invariance (Second-Order)	20872.160 (3634)*	.910	.909	.041	S1	3.629 (3)	.000	.000	.000
S3. Strong invariance (Second-Order)	20968.592 (3637)*	.910	.909	.041	S2	70.538 (3)*	.000	.000	.000
S4. Strict invariance(Second-Order)	20988.002 (3641)*	.910	.909	.041	S3	20.817 (4)*	.000	.000	.000
S5. Latent Var.-Covar. (Second-Order)	20991.534 (3650)*	.910	.909	.041	S4	15.003 (9)	.000	.000	.000
S6. Latent Means (Second-Order)	20992.648 (3651)*	.910	.909	.041	S5	4.142 (1)	.000	.000	.000
Male versus Females Respondents									
F1. Configural invariance (First-Order)	14431.591 (3344)*	.928	.921	.037	--	--	--	--	--
F2. Weak invariance (First-Order)	14465.474 (3434)*	.928	.923	.037	M1	127.191 (90)*	.000	+.002	.000
F3. Strong invariance (First-Order)	14672.288 (3482)*	.927	.923	.037	M2	206.851 (48)*	-.001	.000	.000
F4. Strict invariance(First-Order)	14621.146 (3543)*	.928	.925	.036	M3	114.407 (61)*	+.001	+.002	-.001
F5. Correlated Uniq. (First-Order)	14620.928 (3548)*	.928	.925	.036	M4	7.032 (5)	.000	.000	.000
F6. Latent Var.-Covar. (First-Order)	14768.598 (3606)*	.927	.926	.036	M5	140.614 (58)*	-.001	+.001	.000
F7. Latent Means (First-Order)	14972.673 (3619)*	.926	.925	.036	M6	173.995 (13)*	-.001	-.001	.000
S1. Configural invariance (Second-Order)	15711.451 (3631)*	.921	.921	.037	--	--	--	--	--
S2. Weak invariance (Second-Order)	15729.312 (3634)*	.921	.921	.037	S1	17.193 (3)*	.000	.000	.000
S3. Strong invariance (Second-Order)	15834.875 (3637)*	.920	.920	.037	S2	110.294 (3)*	-.001	-.001	.000
S4. Strict invariance(Second-Order)	15851.223 (3641)*	.920	.920	.037	S3	16.437 (4)*	.000	.000	.000
S5. Latent Var.-Covar. (Second-Order)	15865.988 (3650)*	.920	.920	.037	S4	19.540 (9)	.000	.000	.000
S6. Latent Means (Second-Order)	15865.383 (3651)*	.920	.920	.037	S5	.365 (1)	.000	.000	.000

Description	χ^2 (*df*)	CFI	TLI	RMSEA	CM	$\Delta\chi^2$ (*df*)	ΔCFI	ΔTLI	ΔRMSEA
Military versus Civilian Respondents									
F1. Configural invariance (First-Order)	15290.062 (3344)*	.928	.921	.038	--	--	--	--	--
F2. Weak invariance (First-Order)	15369.004 (3434)*	.928	.923	.038	M1	157.448 (90)*	.000	+.002	.000
F3. Strong invariance (First-Order)	15680.784 (3482)*	.926	.922	.038	M2	296.913 (48)*	-.002	-.001	.000
F4. Strict invariance(First-Order)	15558.166 (3543)*	.927	.925	.037	M3	116.688 (61)*	+.001	+.003	-.001
F5. Correlated Uniq. (First-Order)	15548.809 (3548)*	.927	.925	.037	M4	1.541 (5)	.000	.000	.000
F6. Latent Var.-Covar. (First-Order)	15701.526 (3606)*	.927	.926	.037	M5	146.934 (58)*	.000	+.001	.000
F7. Latent Means (First-Order)	16037.811 (3619)*	.925	.924	.037	M6	271.849 (13)*	-.002	-.002	.000
S1. Configural invariance (Second-Order)	16572.644 (3631)*	.921	.921	.038	--	--	--	--	--
S2. Weak invariance (Second-Order)	16655.270 (3634)*	.921	.920	.038	S1	508.464 (3)*	.000	-.001	.000
S3. Strong invariance (Second-Order)	16918.697 (3637)*	.919	.919	.038	S2	207.491 (3)*	-.002	-.001	.000
S4. Strict invariance(Second-Order)	16944.299 (3641)*	.919	.919	.039	S3	24.798 (4)*	.000	.000	.001
S5. Latent Var.-Covar. (Second-Order)	16954.783 (3650)*	.919	.919	.038	S4	15.689 (9)	.000	.000	-.001
S6. Latent Means (Second-Order)	16952.828 (3651)*	.919	.919	.038	S5	.933 (1)	.000	.000	.000

Note. * $p < .01$; χ^2: robust chi-square test of exact fit; *df*: degrees of freedom; CFI: comparative fit index; TLI: Tucker-Lewis index; RMSEA: root mean square error of approximation; 90% CI: 90% confidence interval; CM: comparison model; Δ: change in fit relative to the CM.

Section 5

Annotated Syntax Files for the Turnover Intentions Measurement Models

For all syntax files:
Mplus commands are in bold.
All command lines in Mplus should end with a semicolon (;)
The title of all command sections (TITLE, DATA, etc.) end with a colon (:).
Text following an exclamation point (!) are annotation and will not be read by Mplus.
Annotations are in greyscale.
These illustrations have been prepared so as to build on one another.

Single level Model (complete input)

TITLE: Single level TI model**;**
!!! The title section allows you to give a title to each of your input files.
DATA: file is data.csv**;**
!!! The data section indicates the name of your data file, which should ideally be in .dat or .csv format.
!!! However, more information on acceptable data format is provided in the Mplus manual.
!!! As long as the data file is in the same folder as the input file, this is sufficient.
!!! If the data file is stored in a separate folder, then a complete path is required.
!!! e.g., D:\Mplus files\data.csv.
VARIABLE:
NAMES ARE ID unit qlang sex status rank bo1 bo2 bo3 bo4 bo5 bo6 bo7 bo8 comp1 comp2 comp3 comp4 aut1 aut2 aut3 aut4 aut5 aut6 pow1 pow2 pow3 Mng1 Mng2 Mng3 k1 k2 k3 k4 k5 k6 k7 k8 k9 k10 ti1 ti2 ti3 mor1 mor2 mor3 mor4 mor5 mor6 je1 je2 je3 je4 je5 je6 je7 je8 je9 je10 je11 je12 je13 je14 je15 je16 je17 je18**;**
!!! The variable section describes your data
!!! The variables included in your data set are named, in order of appearance, after NAMES ARE.
USEVARIABLES ARE ti1 ti2 ti3 **;**
!!!Usevariable lists the variables used in a specific analysis.
MISSING ARE ALL (999)**;**
!!!Indicate the code used to indicate missing responses.
CLUSTER IS unit**;**
!!!Indicate the variable used to identify the higher level unit (here, work units).
IDVARIABLE = ID**;**
!!!Indicate the variable used to uniquely identify each participant.
ANALYSIS:
ESTIMATOR IS MLR; TYPE IS COMPLEX;
!!! The analysis section is used to specific the technical aspects of the analyses.
!!! Here, a Maximum Likelihood estimator robust to non-normality and nesting is requested.
!!! COMPLEX provides a way to control for participants nesting in work units in a single level model.
MODEL:
!!! The model section describes the analytic model
TURN **BY** ti1**@1** ti2 ti3**;** TURN***;**
!!! Here, a single factor, that we name TURN, is estimated (using BY) from items ti1, ti2, and ti3.
!!! To set the scale, one can fix (@) the first loading to 1 and free (*) the factor variance (as above).
!!! Or one could freely estimate all loadings and fixe the variance to 1, as here:
!!! TURN **BY** ti1***** ti2 ti3**;** TURN**@1;**
!!! In single level model, both options are interchangeable.
!!! In multilevel model, the first option (fixing the first loading at 1) is necessary to obtain
!!! a proper variance decomposition across levels.
OUTPUT:
SAMPSTAT STANDARDIZED RESIDUAL CINTERVAL MODINDICES (3.0);
TECH1 TECH2 TECH3 TECH4 SVALUES;
!!! Here is the standard syntax that we use to request various output sections.

Multilevel Model (only sections that differ from the previous one shown)

```
TITLE: Multilevel TI model;
!!! […]
!!! Here, all variables are used at both level. ]
!!! The following would be used to variables only used at L1:
!!! WITHIN = var1 var2;
!!! The following would be used to variables only used at L2:
!!! BETWEEN = var3 var4;
DEFINE:
STANDARDIZE ti1 ti2 ti3 ;
!!! The define section is used to request the standardization of all indicators, to simply interpretations.
!!! This follows recommendations by Marsh et al. (2012) and Morin et al. (2014).
!!! In doubly latent models, all variables used at both levels are automatically group-mean centered.
!!! For variables used only at one level (1 or 2), grand-mean centering is requested by:
!!! CENTER var1 var3 var4 (GRANDMEAN);
!!! For variables used only at level 1, group-mean centering is requested by:
!!! CENTER var2 (GROUPMEAN);
ANALYSIS:
ESTIMATOR IS MLR; TYPE IS TWOLEVEL;
!!! TWOLEVEL requests a multilevel operationalisation.
MODEL:
%WITHIN%
TURN_W BY ti1@1 ti2 ti3;
TURN_W*;
%BETWEEN%
TURN_B BY ti1@1 ti2 ti3;
TURN_B*;
!!! The %WITHIN% section describes the Level 1 model
!!! The %BETWEEN% describes the Level 2 model.
!!! Here, the same factor model is requested at both levels.
!!! The name of the factors (TURN) has been adapted to refer to Level 1 (_W) or 2 (_B)
!!! […]
```

Multilevel Model with Isomorphism (only sections that differ from the previous one shown)

```
TITLE: Multilevel TI model with isomorphism;
!!! […]
MODEL:
%WITHIN%
TURN_W BY ti1@1
ti2 ti3 (t2-t3);
TURN_W*;
%BETWEEN%
TURN_B BY ti1@1
ti2 ti3 (t2-t3);
TURN_B*;
!!! Labels, placed in parentheses, indicate that the loadings associated with the same label are equal..
!!! This is not required for the first loading, already fixed to 1 (and equal) at both levels.
```

Section 6

Annotated Syntax Files for the Psychological Empowerment Measurement Models

Single Level Correlated Factors Model
(only sections differing from those presented in the previous section are shown)

```
TITLE: Single level PE correlated factors model;
!!! […]
ANALYSIS:
ESTIMATOR IS MLR; TYPE IS COMPLEX;
MODEL:
COMP BY comp1@1 comp2 comp3 comp4;
AUT BY aut1@1 aut2 aut3 aut4 aut5 aut6;
POW BY pow1@1 pow2 pow3;
MNG BY Mng1@1 Mng2 Mng3;
COMP*; AUT*; POW*; MNG*;
!!! Four factors (COMP, AUT, POW, MNG) are estimated from their indicators
!!! using specification similar to those described before.
!!! Factor correlations are freely estimated by default.
aut2 WITH aut3 aut6;
aut3 WITH aut6;
!!! a priori correlated uniquenesses are specified (using WITH) to control for the negative wording of
!!! three of the self-determination items (Marsh, Scalas, & Nagengast, 2010).
!!! […]
```

Single Level Higher-Order Model
(only sections differing from those presented in the previous section are shown)

```
TITLE: Single level PE higher-order model;
!!! […]
ANALYSIS:
ESTIMATOR IS MLR; TYPE IS COMPLEX;
MODEL:
COMP BY comp1@1 comp2 comp3 comp4;
AUT BY aut1@1 aut2 aut3 aut4 aut5 aut6;
POW BY pow1@1 pow2 pow3;
MNG BY Mng1@1 Mng2 Mng3;
COMP*; AUT*; POW*; MNG*;
HPE BY COMP@1 AUT POW MNG;
HPE*;
!!! The four factors (COMP, AUT, POW, MNG) are used to estimate a higher-order factor (HPE).
!!! using specification similar to those described before.
!!! Factor correlations are automatically 0 among first-order factors.
aut2 WITH aut3 aut6;
aut3 WITH aut6;
!!! […]
```

Single Level Bifactor Model
(only sections differing from those presented in the previous section are shown)

```
TITLE: Single level PE bifactor model;
!!! […]
ANALYSIS:
ESTIMATOR IS MLR; TYPE IS COMPLEX;
MODEL:
GPE BY comp2@1 comp1 comp3 comp4 aut1 aut2 aut3 aut4 aut5 aut6
pow1 pow2 pow3 Mng1 Mng2 Mng3;
!!! Here a global factor (GPE) is defined from all items
!!! using specification similar to those described before
!!! It is, however, important not the use the same referent indicator for more than one factor
!!! i.e., here, comp1 has a loading fixed to 1 on the COMP factor but not the GPE factor (comp2)
COMP BY comp1@1 comp2 comp3 comp4;
AUT BY aut1@1 aut2 aut3 aut4 aut5 aut6;
POW BY pow1@1 pow2 pow3;
MNG BY Mng1@1 Mng2 Mng3;
COMP*; AUT*; POW*; MNG*;
aut2 WITH aut3 aut6; aut3 WITH aut6;
GPE WITH COMP@0 AUT@0 POW@0 MNG@0;
COMP WITH AUT@0 POW@0 MNG@0;
AUT WITH POW@0 MNG@0;
POW WITH MNG@0;
!!! In bifactor models, the orthogonality of the specific factors has to be specified (correlations @0).
!!! […]
```

Multilevel Level Higher-Order Model
(only sections differing from those presented in the previous section are shown)

```
TITLE: Multilevel higher-order PE model;
DEFINE:
STANDARDIZE comp1 comp2 comp3 comp4 aut1 aut2 aut3 aut4 aut5 aut6;
ANALYSIS:
ESTIMATOR IS MLR; TYPE IS TWOLEVEL;
MODEL:
%WITHIN%
COMP_W BY comp1@1 comp2 comp3 comp4;
AUT_W BY aut1@1 aut2 aut3 aut4 aut5 aut6;
POW_W BY pow1@1 pow2 pow3;
MNG_W BY Mng1@1 Mng2 Mng3;
COMP_W *; AUT_W *; POW_W *; MNG_W *;
HPE_W BY COMP_W @1 AUT_W POW_W MNG_W;
HPE_W *;
aut2 WITH aut3 aut6; aut3 WITH aut6;
%BETWEEN%
COMP_B BY comp1@1 comp2 comp3 comp4;
AUT_B BY aut1@1 aut2 aut3 aut4 aut5 aut6;
POW_B BY pow1@1 pow2 pow3;
MNG_B BY Mng1@1 Mng2 Mng3;
COMP_B *; AUT_B *; POW_B *; MNG_B *;
HPE_B BY COMP_B @1 AUT_B POW_B MNG_B;
HPE_B *;
!!! As for turnover intention, %WITHIN% refers to Level 1 and %BETWEEN% to level 2
!!! As for turnover intention, the same factor model is requested at both levels.
!!! As for turnover intention, the name of the factors refers to Level 1 (_W) or 2 (_B)
```

Multilevel Level Higher-Order Model with Isomorphism
(only sections differing from those presented in the previous section are shown)

```
TITLE: Multilevel higher-order PE model with isomorphism;
!!! […]
MODEL:
%WITHIN%
COMP_W BY comp1@1
comp2 comp3 comp4 (c2-c4);
AUT_W BY aut1@1
aut2 aut3 aut4 aut5 aut6 (a2-a6);
POW_W BY pow1@1
pow2 pow3 (p2-p3);
MNG_W BY Mng1@1
Mng2 Mng3 (m2-m3);
COMP_W *; AUT_W *; POW_W *; MNG_W *;
HPE_W BY COMP_W@1
AUT_W POW_W MNG_W (h2-h4);
HPE_W *;
aut2 WITH aut3 aut6; aut3 WITH aut6;
%BETWEEN%
COMP_B BY comp1@1
comp2 comp3 comp4 (c2-c4);
AUT_B BY aut1@1
aut2 aut3 aut4 aut5 aut6 (a2-a6);
POW_B BY pow1@1
pow2 pow3 (p2-p3);
MNG_B BY Mng1@1
Mng2 Mng3 (m2-m3);
COMP_B*; AUT_B*; POW_B*; MNG_B*;
HPE_B BY COMP_B@1
AUT_B POW_B MNG_B (h2-h4);
HPE_B *;
!!! As for turnover intention, labels (in parentheses), are use to request equal loadings across levels for
!!! all first-order factors and for the higher-order factor.
```

Section 7

Annotated Syntax Files for the Psychological Health Measurement Models

Single Level Correlated Factors Model
(only sections differing from those presented in the previous section are shown)

```
TITLE: Single level PH correlated factors model;
!!! […]
ANALYSIS:
ESTIMATOR IS MLR; TYPE IS COMPLEX;
MODEL:
MOR BY MOR1@1 MOR2 MOR3 MOR4 MOR5 MOR6;
JEP BY JE1@1 JE2 JE3 JE4 JE5 JE6;
JEE BY JE7@1 JE8 JE9 JE10 JE11 JE12;
JEC BY JE13@1 JE14 JE15 JE16 JE17 JE18;
EE BY BO1@1 BO2 BO3 BO4;
DIS BY BO5@1 BO6 BO7 BO8;
PD BY K1@1 K2 K3 K4 K5 K6 K7 K8 K9 K10;
MOR*;
JEP*;
JEE*;
JEC*;
EE*;
DIS*;
PD*;
!!! Seven factors (MOR, JEP, JEE, JEC, EE, DIS, PD) are estimated from their indicators
!!! using specification similar to those described before.
!!! Factor correlations are freely estimated by default.
K2 WITH K3;
K5 WITH K6;
!!! a priori correlated uniquenesses are specified (using WITH) to control for the parallel wording of
!!! two pairs of psychological distress items (Marsh et al., 2013).
!!! […]
```

Single Level Bifactor Model
(only sections differing from those presented in the previous section are shown)

```
TITLE: Single level PH bifactor model;
!!! […]
MODEL:
GPH BY MOR2@1 MOR1 MOR3 MOR4 MOR5 MOR6
JE1 JE2 JE3 JE4 JE5 JE6 JE7 JE8 JE9 JE10 JE11 JE12
JE13 JE14 JE15 JE16 JE17 JE18
BO1 BO2 BO3 BO4 BO5 BO6 BO7 BO8
K1 K2 K3 K4 K5 K6 K7 K8 K9 K10 ;
GPH*;
!!! Here a global factor (GPH) is defined from all items
!!! using specification similar to those described before.
!!! It is, however, important not the use the same referent indicator for more than one factor
!!! i.e., here, mor1 has a loading fixed to 1 on the MOR factor but not the GPE factor (mor2)
MOR BY MOR1@1 MOR2 MOR3 MOR4 MOR5 MOR6;
JEP BY JE1@1 JE2 JE3 JE4 JE5 JE6;
JEE BY JE7@1 JE8 JE9 JE10 JE11 JE12;
JEC BY JE13@1 JE14 JE15 JE16 JE17 JE18;
EE BY BO1@1 BO2 BO3 BO4;
DIS BY BO5@1 BO6 BO7 BO8;
PD BY K1@1 K2 K3 K4 K5 K6 K7 K8 K9 K10;
MOR*;
JEP*;
JEE*;
JEC*;
EE*;
DIS*;
PD*;
K2 WITH K3;
K5 WITH K6;
GPH WITH MOR@0 JEP@0 JEE@0 JEC@0 EE@0 DIS@0 PD@0;
MOR WITH JEP@0 JEE@0 JEC@0 EE@0 DIS@0 PD@0;
JEP WITH JEE@0 JEC@0 EE@0 DIS@0 PD@0;
JEE WITH JEC@0 EE@0 DIS@0 PD@0;
JEC WITH EE@0 DIS@0 PD@0;
EE WITH DIS@0 PD@0;
DIS WITH PD@0
!!! In bifactor models, the orthogonality of the specific factors has to be specified (correlations @0).
!!! […]
```

Multilevel Level Bifactor Model
(only sections differing from those presented in the previous section are shown)

```
TITLE: Multilevel bifactor PH model;
!!! […]
DEFINE:
STANDARDIZE bo1 bo2 bo3 bo4 bo5 bo6 bo7 bo8 k1 k2 k3 k4 k5 k6 k7 k8 k9 k10
MOR1 MOR2 MOR3 MOR4 MOR5 MOR6 je1 je2 je3 je4 je5 je6 je7 je8 je9 je10 je11
je12 je13 je14 je15 je16 je17 je18;
ANALYSIS:
ESTIMATOR IS MLR; TYPE IS TWOLEVEL;
MODEL:
%WITHIN%
GPH_W BY MOR2@1 MOR1 MOR3 MOR4 MOR5 MOR6 JE1 JE2 JE3 JE4 JE5 JE6 JE7
JE8 JE9 JE10 JE11 JE12 JE13 JE14 JE15 JE16 JE17 JE18 BO1 BO2 BO3 BO4 BO5 BO6
BO7 BO8 K1 K2 K3 K4 K5 K6 K7 K8 K9 K10 ;
GPH_W*;
MOR_W BY MOR1@1 MOR2 MOR3 MOR4 MOR5 MOR6;
JEP_W BY JE1@1 JE2 JE3 JE4 JE5 JE6;
JEE_W BY JE7@1 JE8 JE9 JE10 JE11 JE12;
JEC_W BY JE13@1 JE14 JE15 JE16 JE17 JE18;
EE_W BY BO1@1 BO2 BO3 BO4;
DIS_W BY BO5@1 BO6 BO7 BO8;
PD_W BY K1@1 K2 K3 K4 K5 K6 K7 K8 K9 K10;
MOR_W*; JEP_W*; JEE_W*; JEC_W*; EE_W*; DIS_W*; PD_W*;
K2 WITH K3; K5 WITH K6;
GPH_W WITH MOR_W@0 JEP_W@0 JEE_W@0 JEC_W@0 EE_W@0 DIS_W@0 PD_W@0;
MOR_W WITH JEP_W@0 JEE_W@0 JEC_W@0 EE_W@0 DIS_W@0 PD_W@0;
JEP_W WITH JEE_W@0 JEC_W@0 EE_W@0 DIS_W@0 PD_W@0;
JEE_W WITH JEC_W@0 EE_W@0 DIS_W@0 PD_W@0;
JEC_W WITH EE_W@0 DIS_W@0 PD_W@0;
EE_W WITH DIS_W@0 PD_W@0;
DIS_W WITH PD_W@0;
%BETWEEN%
GPH_B BY MOR2@1 MOR1 MOR3 MOR4 MOR5 MOR6 JE1 JE2 JE3 JE4 JE5 JE6 JE7
JE8 JE9 JE10 JE11 JE12 JE13 JE14 JE15 JE16 JE17 JE18 BO1 BO2 BO3 BO4 BO5 BO6
BO7 BO8 K1 K2 K3 K4 K5 K6 K7 K8 K9 K10 ;
GPH_B*;
MOR_B BY MOR1@1 MOR2 MOR3 MOR4 MOR5 MOR6;
JEP_B BY JE1@1 JE2 JE3 JE4 JE5 JE6;
JEE_B BY JE7@1 JE8 JE9 JE10 JE11 JE12;
JEC_B BY JE13@1 JE14 JE15 JE16 JE17 JE18;
EE_B BY BO1@1 BO2 BO3 BO4;
DIS_B BY BO5@1 BO6 BO7 BO8;
PD_B BY K1@1 K2 K3 K4 K5 K6 K7 K8 K9 K10;
MOR_B*; JEP_B*; JEE_B*; JEC_B*; EE_B*; DIS_B*; PD_B*;
GPH_B WITH MOR_B@0 JEP_B@0 JEE_B@0 JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
MOR_B WITH JEP_B@0 JEE_B@0 JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
JEP_B WITH JEE_B@0 JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
JEE_B WITH JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
JEC_B WITH EE_B@0 DIS_B@0 PD_B@0;
EE_B WITH DIS_B@0 PD_B@0;
DIS_B WITH PD_B@0;
!!! As for turnover intention, %WITHIN% refers to Level 1 and %BETWEEN% to level 2
!!! As for turnover intention, the same factor model is requested at both levels.
!!! As for turnover intention, the name of the factors refers to Level 1 (_W) or 2 (_B)
!!! […]
```

Multilevel Level Bifactor Model with Isomorphism
(only input sections differing from those presented in the previous section are shown)

TITLE: Multilevel bifactor PH model with isomorphism**;**
!!! **[…]**
MODEL:
%WITHIN%
GPH_W **BY** MOR2**@1;**
GPH_W **BY** MOR1***** MOR3 MOR4 MOR5 MOR6 (L36-L40)**;**
GPH_W **BY** JE1***** JE2 JE3 JE4 JE5 JE6 (L41-L46)**;**
GPH_W **BY** JE7***** JE8 JE9 JE10 JE11 JE12 (L47-L52)**;**
GPH_W **BY** JE13***** JE14 JE15 JE16 JE17 JE18 (L53-L58)**;**
GPH_W **BY** BO1***** BO2 BO3 BO4 (L59-L62)**;**
GPH_W **BY** BO5***** BO6 BO7 BO8 (L63-L66)**;**
GPH_W **BY** K1***** K2 K3 K4 K5 K6 K7 K8 K9 K10 (L67-L76)**;**
GPH_W***;**
MOR_W **BY** MOR1**@1**
MOR2 MOR3 MOR4 MOR5 MOR6 (L1-L5)**;**
JEP_W **BY** JE1**@1**
JE2 JE3 JE4 JE5 JE6 (L6-L10)**;**
JEE_W **BY** JE7**@1**
JE8 JE9 JE10 JE11 JE12 (L11-L15)**;**
JEC_W **BY** JE13**@1**
JE14 JE15 JE16 JE17 JE18 (L16-L20)**;**
EE_W **BY** BO1**@1**
BO2 BO3 BO4 (L21-L23)**;**
DIS_W **BY** BO5**@1**
BO6 BO7 BO8 (L24-L26)**;**
PD_W **BY** K1**@1**
K2 K3 K4 K5 K6 K7 K8 K9 K10 (L27-L35)**;**
MOR_W***;** JEP_W***;** JEE_W***;** JEC_W***;** EE_W***;** DIS_W***;** PD_W***;**
K2 **WITH** K3**;** K5 **WITH** K6**;**
GPH_W **WITH** MOR_W**@0** JEP_W**@0** JEE_W**@0** JEC_W**@0** EE_W**@0** DIS_W**@0** PD_W**@0;**
MOR_W **WITH** JEP_W**@0** JEE_W**@0** JEC_W**@0** EE_W**@0** DIS_W**@0** PD_W**@0;**
JEP_W **WITH** JEE_W**@0** JEC_W**@0** EE_W**@0** DIS_W**@0** PD_W**@0;**
JEE_W **WITH** JEC_W**@0** EE_W**@0** DIS_W**@0** PD_W**@0;**
JEC_W **WITH** EE_W**@0** DIS_W**@0** PD_W**@0;**
EE_W **WITH** DIS_W**@0** PD_W**@0;**
DIS_W **WITH** PD_W**@0;**
%BETWEEN%
GPH_B **BY** MOR2**@1;**
GPH_B **BY** MOR1***** MOR3 MOR4 MOR5 MOR6 (L36-L40)**;**
GPH_B **BY** JE1***** JE2 JE3 JE4 JE5 JE6 (L41-L46)**;**
GPH_B **BY** JE7***** JE8 JE9 JE10 JE11 JE12 (L47-L52)**;**
GPH_B **BY** JE13***** JE14 JE15 JE16 JE17 JE18 (L53-L58)**;**
GPH_B **BY** BO1***** BO2 BO3 BO4 (L59-L62)**;**
GPH_B **BY** BO5***** BO6 BO7 BO8 (L63-L66)**;**
GPH_B **BY** K1***** K2 K3 K4 K5 K6 K7 K8 K9 K10 (L67-L76)**;**
GPH_B***;**
MOR_B **BY** MOR1**@1**
MOR2 MOR3 MOR4 MOR5 MOR6 (L1-L5)**;**
JEP_B **BY** JE1**@1**
JE2 JE3 JE4 JE5 JE6 (L6-L10)**;**
JEE_B **BY** JE7**@1**
JE8 JE9 JE10 JE11 JE12 (L11-L15)**;**

```
JEC_B BY JE13@1
JE14 JE15 JE16 JE17 JE18 (L16-L20);
EE_B BY BO1@1
BO2 BO3 BO4 (L21-L23);
DIS_B BY BO5@1
BO6 BO7 BO8 (L24-L26);
PD_B BY K1@1
K2 K3 K4 K5 K6 K7 K8 K9 K10 (L27-L35);
MOR_B*; JEP_B*; JEE_B*; JEC_B*; EE_B*; DIS_B*; PD_B*;
GPH_B WITH MOR_B@0 JEP_B@0 JEE_B@0 JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
MOR_B WITH JEP_B@0 JEE_B@0 JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
JEP_B WITH JEE_B@0 JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
JEE_B WITH JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
JEC_B WITH EE_B@0 DIS_B@0 PD_B@0;
EE_B WITH DIS_B@0 PD_B@0;
DIS_B WITH PD_B@0;
!!! As for turnover intention, labels (in parentheses), are use to request equal loadings across levels for
!!! all specific and global factors.
```

Section 8

Syntax File for the Complete Measurement Model (with Isomorphism)

(As this input is simply the combination of previous one, no annotations are required)

```
TITLE: Complete multilevel model with isomorphism;
DATA: file is data.csv;
VARIABLE:
NAMES ARE ID unit qlang sex status rank bo1 bo2 bo3 bo4 bo5 bo6 bo7 bo8 comp1 comp2
comp3 comp4 aut1 aut2 aut3 aut4 aut5 aut6 pow1 pow2 pow3 Mng1 Mng2 Mng3 k1 k2 k3 k4
k5 k6 k7 k8 k9 k10 ti1 ti2 ti3 mor1 mor2 mor3 mor4 mor5 mor6 je1 je2 je3 je4 je5 je6 je7 je8
je9 je10 je11 je12 je13 je14 je15 je16 je17 je18;
USEVARIABLES ARE comp1 comp2 comp3 comp4 aut1 aut2 aut3 aut4 aut5 aut6
pow1 pow2 pow3 Mng1 Mng2 Mng3 ti1 ti2 ti3 bo1 bo2 bo3 bo4 bo5 bo6 bo7 bo8
k1 k2 k3 k4 k5 k6 k7 k8 k9 k10 MOR1 MOR2 MOR3 MOR4 MOR5 MOR6
je1 je2 je3 je4 je5 je6 je7 je8 je9 je10 je11 je12 je13 je14 je15 je16 je17 je18;;
MISSING ARE ALL (999);
CLUSTER IS unit;
IDVARIABLE = ID;
ANALYSIS:
ESTIMATOR IS MLR; TYPE IS TWOLEVEL;
MODEL:
%WITHIN%
!!! Turnover Intentions Level 1
TURN_W BY ti1@1
ti2 ti3 (t2-t3);
TURN_W*;
!!! Psychological Empowerment Level 1
COMP_W BY comp1@1
comp2 comp3 comp4 (c2-c4);
AUT_W BY aut1@1
aut2 aut3 aut4 aut5 aut6 (a2-a6);
POW_W BY pow1@1
pow2 pow3 (p2-p3);
MNG_W BY Mng1@1
Mng2 Mng3 (m2-m3);
COMP_W *; AUT_W *; POW_W *; MNG_W *;
HPE_W BY COMP_W@1
AUT_W POW_W MNG_W (h2-h4);
HPE_W *;
aut2 WITH aut3 aut6; aut3 WITH aut6;
!!! Psychological Health Level 1
GPH_W BY MOR2@1;
GPH_W BY MOR1* MOR3 MOR4 MOR5 MOR6 (L36-L40);
GPH_W BY JE1* JE2 JE3 JE4 JE5 JE6 (L41-L46);
GPH_W BY JE7* JE8 JE9 JE10 JE11 JE12 (L47-L52);
GPH_W BY JE13* JE14 JE15 JE16 JE17 JE18 (L53-L58);
GPH_W BY BO1* BO2 BO3 BO4 (L59-L62);
GPH_W BY BO5* BO6 BO7 BO8 (L63-L66);
GPH_W BY K1* K2 K3 K4 K5 K6 K7 K8 K9 K10 (L67-L76);
GPH_W*;
MOR_W BY MOR1@1
MOR2 MOR3 MOR4 MOR5 MOR6 (L1-L5);
JEP_W BY JE1@1
JE2 JE3 JE4 JE5 JE6 (L6-L10);
```

```
JEE_W BY JE7@1
JE8 JE9 JE10 JE11 JE12 (L11-L15);
JEC_W BY JE13@1
JE14 JE15 JE16 JE17 JE18 (L16-L20);
EE_W BY BO1@1
BO2 BO3 BO4 (L21-L23);
DIS_W BY BO5@1
BO6 BO7 BO8 (L24-L26);
PD_W BY K1@1
K2 K3 K4 K5 K6 K7 K8 K9 K10 (L27-L35);
MOR_W*; JEP_W*; JEE_W*; JEC_W*; EE_W*; DIS_W*; PD_W*;
K2 WITH K3; K5 WITH K6;
GPH_W WITH MOR_W@0 JEP_W@0 JEE_W@0 JEC_W@0 EE_W@0 DIS_W@0 PD_W@0;
MOR_W WITH JEP_W@0 JEE_W@0 JEC_W@0 EE_W@0 DIS_W@0 PD_W@0;
JEP_W WITH JEE_W@0 JEC_W@0 EE_W@0 DIS_W@0 PD_W@0;
JEE_W WITH JEC_W@0 EE_W@0 DIS_W@0 PD_W@0;
JEC_W WITH EE_W@0 DIS_W@0 PD_W@0;
EE_W WITH DIS_W@0 PD_W@0;
DIS_W WITH PD_W@0;
%BETWEEN%
!!! Turnover Intentions Level 2
TURN_B BY ti1@1
ti2 ti3 (t2-t3);
TURN_B*;
!!! Psychological Empowerment Level 2
COMP_B BY comp1@1
comp2 comp3 comp4 (c2-c4);
AUT_B BY aut1@1
aut2 aut3 aut4 aut5 aut6 (a2-a6);
POW_B BY pow1@1
pow2 pow3 (p2-p3);
MNG_B BY Mng1@1
Mng2 Mng3 (m2-m3);
COMP_B*; AUT_B*; POW_B*; MNG_B*;
HPE_B BY COMP_B@1
AUT_B POW_B MNG_B (h2-h4);
HPE_B *;
!!! Psychological Health Level 2
GPH_B BY MOR2@1;
GPH_B BY MOR1* MOR3 MOR4 MOR5 MOR6 (L36-L40);
GPH_B BY JE1* JE2 JE3 JE4 JE5 JE6 (L41-L46);
GPH_B BY JE7* JE8 JE9 JE10 JE11 JE12 (L47-L52);
GPH_B BY JE13* JE14 JE15 JE16 JE17 JE18 (L53-L58);
GPH_B BY BO1* BO2 BO3 BO4 (L59-L62);
GPH_B BY BO5* BO6 BO7 BO8 (L63-L66);
GPH_B BY K1* K2 K3 K4 K5 K6 K7 K8 K9 K10 (L67-L76);
GPH_B*;
MOR_B BY MOR1@1
MOR2 MOR3 MOR4 MOR5 MOR6 (L1-L5);
JEP_B BY JE1@1
JE2 JE3 JE4 JE5 JE6 (L6-L10);
JEE_B BY JE7@1
JE8 JE9 JE10 JE11 JE12 (L11-L15);
JEC_B BY JE13@1
JE14 JE15 JE16 JE17 JE18 (L16-L20);
```

```
EE_B BY BO1@1
BO2 BO3 BO4 (L21-L23);
DIS_B BY BO5@1
BO6 BO7 BO8 (L24-L26);
PD_B BY K1@1
K2 K3 K4 K5 K6 K7 K8 K9 K10 (L27-L35);
MOR_B*; JEP_B*; JEE_B*; JEC_B*; EE_B*; DIS_B*; PD_B*;
GPH_B WITH MOR_B@0 JEP_B@0 JEE_B@0 JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
MOR_B WITH JEP_B@0 JEE_B@0 JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
JEP_B WITH JEE_B@0 JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
JEE_B WITH JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
JEC_B WITH EE_B@0 DIS_B@0 PD_B@0;
EE_B WITH DIS_B@0 PD_B@0;
DIS_B WITH PD_B@0;
OUTPUT:
SAMPSTAT STANDARDIZED RESIDUAL CINTERVAL MODINDICES (3.0);
TECH1 TECH2 TECH3 TECH4 SVALUES;
```

Section 9
Reliability Calculation

Inter-Item Reliability

The first two sources of measurement error, related to "inter-item" agreement in the estimation of latent constructs, are assessed using McDonald' (1970) omega (ω) coefficient of composite reliability. This coefficient is calculated using the standardized factor loadings and uniquenesses from the retained single-level and multilevel measurement models using the following formula.

$$\omega = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \left(\sum_{i=1}^{k} \theta_{ii}\right)}$$

where λ_i reflect the standardized factor loadings associated with the i^{th} item (out of k items) on a specific factor, and θ_{ij} are the standardized item uniquenesses. This coefficient thus provides a direct estimate of the ratio of true score variance on the total variance present at the item level. As this coefficients involved standardized model parameters, it needs to be calculated by hand.

A very similar coefficient, called rho (ρ), has been proposed by Raykov (2009) and can be calculated using the model unstandardized parameters. Because it involves the model unstandardized parameters, its calculation can be automated in Mplus. The calculation of ρ is fairly straightforward (and relies on the same formula used for ω but using the unstandardized parameters) when the scale of the latent factors is set by fixing the factor variance to 1 and freely estimating all factor loadings. However, the calculation of ρ is more complex when the scale of the latent factors is set by fixing the loading of a referent indicator to 1 and freely estimating the factor variance:

$$\rho = \frac{(\sum\lambda)^2\phi}{(\sum\lambda)^2\phi + \sum\delta}$$

where ϕ refers to the factor variance. In practice, we find the manual calculation of ω to be fairly simple, and simpler than the automated calculation of ρ, leading us to recommend focusing on ω (which is also the most commonly reported indicator of composite reliability).

To illustrate the calculation of ω, ω_{L1}, ω_{L2}, we rely on the results associated with the single level and multilevel measurement models of TI. In Mplus, the standardized parameter estimates from the single level measurement model of TI appear in the following section of the output, where the relevant numbers are in bold:

STANDARDIZED MODEL RESULTS
STDYX Standardization

			Two-Tailed	
	Estimate	S.E.	Est./S.E.	P-Value
TURN BY				
TI1	**0.678**	0.014	48.576	0.000
TI2	**0.805**	0.008	103.495	0.000
TI3	**0.863**	0.011	75.259	0.000
Intercepts				
TI1	1.707	0.021	82.918	0.000
TI2	1.851	0.027	69.485	0.000
TI3	1.761	0.025	70.056	0.000
Variances				
TURN	1.000	0.000	999.000	999.000
Residual Variances				
TI1	**0.541**	0.019	28.580	0.000
TI2	**0.352**	0.013	28.084	0.000
TI3	**0.255**	0.020	12.901	0.000

To calculate ω, one first sum the factor loadings (BY: .678+.805+.863 = 2.346) and then to square this sum (2.346^2 = 5.503716). One then has to sum the uniquenesses (Residual Variances: .541+.352+.255 = 1.148) and to sum both numbers (5.503716 + 1.148 = 6.651716). The ω coefficient is obtained by dividing the squared sum of loadings by the final number (5.503716 / 6.651716 = .827).

For the final isomorphic multilevel results, the relevant parameter estimates are found in the following section of the output, where numbers relevant to L1 reliability are indicated in bold, whereas those relevant to L2 reliability are underlined:

STANDARDIZED MODEL RESULTS
STDYX Standardization

			Two-Tailed	
	Estimate	S.E.	Est./S.E.	P-Value
Within Level				
TURN BY				
TI1	**0.674**	0.014	47.694	0.000
TI2	**0.810**	0.008	97.782	0.000
TI3	**0.862**	0.011	75.046	0.000
Variances				
TURN	1.000	0.000	999.000	999.000
Residual Variances				
TI1	**0.546**	0.019	28.693	0.000
TI2	**0.345**	0.013	25.697	0.000
TI3	**0.256**	0.020	12.919	0.000
Between Level				
TURN_B BY				
TI1	<u>0.661</u>	0.087	7.611	0.000
TI2	<u>0.683</u>	0.092	7.449	0.000
TI3	<u>0.927</u>	0.099	9.377	0.000
Intercepts				
TI1	0.005	0.191	0.024	0.981
TI2	-0.088	0.149	-0.586	0.558
TI3	0.083	0.182	0.457	0.648
Variances				
TURN_B	1.000	0.000	999.000	999.000
Residual Variances				
TI1	<u>0.563</u>	0.115	4.905	0.000
TI2	<u>0.534</u>	0.125	4.271	0.000
TI3	<u>0.140</u>	0.183	0.766	0.444

Using the same calculations as above lead to ω_{L1} = .828 (5.503716/ 6.650716) and ω_{L2} = .807 (5.157441 / 6.394441). Alternatively, calculating ρ in a single level model would require using parameter labels for the factor loadings, factor variances, and item uniquenesses, and using these labels in a new MODEL CONSTRAINT section to request the calculation of ρ:

```
MODEL:
TURN BY TI1@1
TI2 TI3 (L2-L3);
TURN* (V1);
TI1-TI3 (I1-I3);
MODEL CONSTRAINT:
NEW (RHO);
RHO = = (1 + L2 + L3)**2* V1/((1 + L2 + L3)**2 * V1 + I1 + I2 + I3);
!!! 1 is used to refer to the first factor loading fixed to 1.
```

Using this set-up, the result will appear in the following section of the output (in bold). Differences between ω (.827) and ρ (.829) are typically minimal.

MODEL RESULTS

			Two-Tailed	
	Estimate	S.E.	Est./S.E.	P-Value
TURN BY				
TI1	1.000	0.000	999.000	999.000
TI2	1.215	0.027	45.245	0.000
TI3	1.321	0.037	35.464	0.000
Intercepts				
TI1	2.381	0.042	56.670	0.000
TI2	2.642	0.042	62.682	0.000
TI3	2.549	0.039	64.649	0.000
Variances				
TURN	0.894	0.042	21.520	0.000
Residual Variances				
TI1	1.052	0.047	22.333	0.000
TI2	0.717	0.028	25.362	0.000
TI3	0.535	0.042	12.707	0.000
New/Additional Parameters				
RHO	**0.829**	0.006	137.902	0.000

In the multilevel model, $\rho_{L1}$and ρ_{L2} will be calculated in the following manner:

```
MODEL:
%WITHIN%
TURN_W BY TI1@1
TI2 TI3 (L2-L3);
TI1-TI3 (I1-I3);
TURN_W* (V1);
%BETWEEN%
TURN_B BY  TI1@1
TI2 TI3 (L2-L3);
TI1-TI3 (IB1-IB3);
TURN_B* (V2);
MODEL CONSTRAINT:
NEW (RHO_L1);
RHO_L1 = (1 + L2 + L3)**2* V1/((1 + L2 + L3)**2 * V1 + I1 + I2 + I3);
NEW (RHO_L2);
RHO_L2 = (1 + L2 + L3)**2* V2/((1 + L2 + L3)**2 * V2 + IB1 + IB2 + IB3);
!!! Here loadings are equal (with the same labels) across levels due to isomorphism.
```

Finally, the results reveal that $\rho_{L1} = .828$ and $\rho_{L2} = .791$ (see below, in bold):

MODEL RESULTS

			Two-Tailed	
	Estimate	S.E.	Est./S.E.	P-Value
Within Level				
TURN_W BY				
TI1	1.000	0.000	999.000	999.000
TI2	1.202	0.027	45.327	0.000
TI3	1.285	0.036	35.354	0.000
Variances				
TURN_W	0.441	0.021	20.916	0.000
Residual Variances				
TI1	0.531	0.023	22.660	0.000
TI2	0.335	0.014	23.591	0.000
TI3	0.251	0.019	13.133	0.000
Between Level				
TURN_B BY				
TI1	1.000	0.000	999.000	999.000
TI2	1.202	0.027	45.327	0.000
TI3	1.285	0.036	35.354	0.000
Intercepts				
TI1	0.001	0.030	0.024	0.981
TI2	-0.016	0.028	-0.585	0.559
TI3	0.012	0.025	0.472	0.637
Variances				
TURN_B	0.011	0.004	2.589	0.010
Residual Variances				
TI1	0.014	0.004	3.331	0.001
TI2	0.018	0.007	2.748	0.006
TI3	0.003	0.004	0.722	0.470
New/Additional Parameters				
RHO_L1	**0.828**	0.006	136.181	0.000
RHO_L2	**0.791**	0.070	11.324	0.000

Inter-Rater Reliability

The final source of measurement error, related to the degree of "inter-rater" agreement in the assessment of the L2 construct can be estimated by considering two intra-class correlation coefficients (ICC1 and ICC2; Marsh et al., 2012). The more common ICC1 indicates the proportion of the total variance in rating occurring at L2 and reflects the average agreement among members of a single L2 unit in ratings of a specific construct (*x*), and is calculated as:

$$ICC1 = \frac{\tau_x^2}{\tau_x^2 + \sigma_x^2}$$

where τ_x^2 refers to the L2 variance, σ_x^2 to the L1 variance. In contrast, the ICC2 reflects the reliability of the group (L2) aggregate, and is calculated as:

$$ICC2 = \frac{\tau_x^2}{\tau_x^2 + \left(\sigma_x^2 / n_j\right)}$$

where τ_x^2 refers to the L2 variance, σ_x^2 to the L1 variance, and n_j to the average number of participants in each of the L2 units. Once again, our suggestion is to calculate these by hand, using the parameter estimates from the retained multilevel model. We again rely on the results obtained from the simpler TI model for illustration purposes. In the isomorphic multilevel TI measurement model, the L1 (σ_x^2: in bold) and L2 (τ_x^2: underlined) variance estimates can be found in the following section:

MODEL RESULTS

	Estimate	S.E.	Two-Tailed Est./S.E.	P-Value
Within Level				
TURN BY				
TI1	1.000	0.000	999.000	999.000
TI2	1.202	0.027	45.327	0.000
TI3	1.285	0.036	35.354	0.000
Variances				
TURN	**0.441**	0.021	20.916	0.000
Residual Variances				
TI1	0.531	0.023	22.660	0.000
TI2	0.335	0.014	23.591	0.000
TI3	0.251	0.019	13.133	0.000
Between Level				
TURN_B BY				
TI1	1.000	0.000	999.000	999.000
TI2	1.202	0.027	45.327	0.000
TI3	1.285	0.036	35.354	0.000
Intercepts				
TI1	0.001	0.030	0.024	0.981
TI2	-0.016	0.028	-0.585	0.559
TI3	0.012	0.025	0.472	0.637
Variances				
TURN_B	<u>0.011</u>	0.004	2.589	0.010
Residual Variances				
TI1	0.014	0.004	3.331	0.001
TI2	0.018	0.007	2.748	0.006
TI3	0.003	0.004	0.722	0.470

Knowing that the average cluster size is 119.898 (n_j), the ICC1 can be calculated as .011 / (.011 + .441) = .024, whereas the ICC2 can be calculated as .011 / (.011 + (.441/119.898)) = .749. Once again, however, it is possible to automate these relatively simple calculations using the MODEL CONSTRAINT function (Raykov, 2011). In the input, one simply has to label the variance estimates:

```
MODEL:
%WITHIN%
TURN_W BY TI1@1
TI2 TI3 (L2-L3);
TI1-TI3 (I1-I3);
TURN_W* (V1);
%BETWEEN%
TURN_B BY  TI1@1
TI2 TI3 (L2-L3);
TI1-TI3 (IB1-IB3);
TURN_B* (V2);
MODEL CONSTRAINT:
NEW (ICC1);  ICC1 = V2 /(V1 + V2);
NEW (ICC2);  ICC2 = V2 /(V1/119.898 + V2);
!!! Here loadings are equal (with the same labels) across levels due to isomorphism.
```

The results (in bold) replicate our manual calculations, although slight differences are possible due to rounding, as shown in the following output section:

```
MODEL RESULTS
Within Level
 TURN_W   BY
   TI1            1.000      0.000    999.000    999.000
   TI2            1.202      0.027     45.327      0.000
   TI3            1.285      0.036     35.354      0.000
 Variances
   TURN_W         0.441      0.021     20.916      0.000
 Residual Variances
   TI1            0.531      0.023     22.660      0.000
   TI2            0.335      0.014     23.591      0.000
   TI3            0.251      0.019     13.133      0.000
Between Level
 TURN_B   BY
   TI1            1.000      0.000    999.000    999.000
   TI2            1.202      0.027     45.327      0.000
   TI3            1.285      0.036     35.354      0.000
 Intercepts
   TI1            0.001      0.030      0.024      0.981
   TI2           -0.016      0.028     -0.585      0.559
   TI3            0.012      0.025      0.472      0.637
 Variances
   TURN_B         0.011      0.004      2.589      0.010
 Residual Variances
   TI1            0.014      0.004      3.331      0.001
   TI2            0.018      0.007      2.748      0.006
   TI3            0.003      0.004      0.722      0.470
New/Additional Parameters
   ICC1           0.024      0.009      2.611      0.009
   ICC2           0.749      0.074     10.122      0.000
```

References

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S. & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47,* 106-124.

McDonald, R.P. (1970). Theoretical foundations of principal factor analysis and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology, 23,* 1-21.

Raykov, T. (2009). Evaluation of scale reliability for unidimensional measures using latent variable modeling. *Measurement & Evaluation in Counselling & Development, 42,* 223-232.

Raykov, T. (2011). Interclass correlation coefficients in hierarchical designs: Evaluation using latent variable modeling. *Structural Equation Modeling, 18,* 73-90.

Section 10
Syntax File for the Predictive DL-MLSEM Models

Essentially, this syntax is identical to that presented in Section 8 of these supplements in relation to the measurement part of the models. We thus only indicate the commands that need to be added in the %WITHIN% and %BETWEEN% sections of the input to convert these models to predictive ones.

M1. Multilevel Predictive: A priori model

```
!!! […]
MODEL:
%WITHIN%
!!! […]
GPH_W WITH COMP_W@0 AUT_W@0 POW_W@0 MNG_W@0 ;
MOR_W WITH COMP_W@0 AUT_W@0 POW_W@0 MNG_W@0 HPE_W@0;
JEP_W WITH COMP_W@0 AUT_W@0 POW_W@0 MNG_W@0 HPE_W@0;
JEE_W WITH COMP_W@0 AUT_W@0 POW_W@0 MNG_W@0 HPE_W@0;
JEC_W WITH COMP_W@0 AUT_W@0 POW_W@0 MNG_W@0 HPE_W@0;
EE_W WITH COMP_W@0 AUT_W@0 POW_W@0 MNG_W@0 HPE_W@0;
DIS_W WITH COMP_W@0 AUT_W@0 POW_W@0 MNG_W@0 HPE_W@0;
!!! The above commands are to specify that the first-order PE factors are unrelated to all PH factors
!!! and that the higher-order PE factor is also unrelated to the specific PH factors.
TURN_W ON GPH_W (TG_W);
TURN_W ON HPE_W (TH_W);
GPH_W ON HPE_W (GH_W);
!!! These are the main predictive paths, and each of them is labelled (in parenthesis).
!!! Predictions are specified with the ON command.
%BETWEEN%
!!! […]
GPH_B WITH COMP_B@0 AUT_B@0 POW_B@0 MNG_B@0 ;
MOR_B WITH COMP_B@0 AUT_B@0 POW_B@0 MNG_B@0 HPE_B@0;
JEP_B WITH COMP_B@0 AUT_B@0 POW_B@0 MNG_B@0 HPE_B@0;
JEE_B WITH COMP_B@0 AUT_B@0 POW_B@0 MNG_B@0 HPE_B@0;
JEC_B WITH COMP_B@0 AUT_B@0 POW_B@0 MNG_B@0 HPE_B@0;
EE_B WITH COMP_B@0 AUT_B@0 POW_B@0 MNG_B@0 HPE_B@0;
DIS_B WITH COMP_B@0 AUT_B@0 POW_B@0 MNG_B@0 HPE_B@0;
TURN_B ON GPH_B (TG_B);
TURN_B ON HPE_B (TH_B);
GPH_B ON HPE_B (GH_B);
!!! The same model is requested in the %BETWEEN% section.
```

M2. M1 + Specific PE —> TI

```
!!! […]
MODEL:
%WITHIN%
!!! […]
TURN_W ON GPH_W (TG_W);
TURN_W ON HPE_W (TH_W);
GPH_W ON HPE_W (GH_W);
!!! Same as before.
HPE_W WITH MOR_W@0 JEP_W@0 JEE_W@0 JEC_W@0 EE_W@0 DIS_W@0 PD_W@0;
GPH_W WITH COMP_W@0 AUT_W@0 POW_W@0 MNG_W@0;
!!! These are used to replace the previous correlations fixed to 0, to indicate a lack of relation between
!!! the PE higher order factor and he specific PH factor, as well as between the PH global factor and
!!! the PE specific factors.
 TURN_W ON COMP_W AUT_W POW_W MNG_W;
!!! These are the additional Predictive paths.
%BETWEEN%
!!! […]
TURN_B ON GPH_B (TG_B);
TURN_B ON HPE_B (TH_B);
GPH_B ON HPE_B (GH_B);
HPE_B WITH MOR_B@0 JEP_B@0 JEE_B@0 JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
GPH_B WITH COMP_B@0 AUT_B@0 POW_B@0 MNG_B@0;
TURN_B ON COMP_B AUT_B POW_B MNG_B;
!!! The same model is requested in the %BETWEEN% section.
```

M3. M1 + Specific PH —> TI (change from the previous is marked in greyscale)

```
!!! […]
MODEL:
%WITHIN%
!!! […]
TURN_W ON GPH_W (TG_W);
TURN_W ON HPE_W (TH_W);
GPH_W ON HPE_W (GH_W);
HPE_W WITH MOR_W@0 JEP_W@0 JEE_W@0 JEC_W@0 EE_W@0 DIS_W@0 PD_W@0;
GPH_W WITH COMP_W@0 AUT_W@0 POW_W@0 MNG_W@0;
TURN_W ON MOR_W JEP_W JEE_W JEC_W EE_W DIS_W PD_W;
%BETWEEN%
!!! […]
TURN_B ON GPH_B (TG_B);
TURN_B ON HPE_B (TH_B);
GPH_B ON HPE_B (GH_B);
HPE_B WITH MOR_B@0 JEP_B@0 JEE_B@0 JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
GPH_B WITH COMP_B@0 AUT_B@0 POW_B@0 MNG_B@0;
TURN_B ON MOR_B JEP_B JEE_B JEC_B EE_B DIS_B PD_B;
```

M4. M1 + Specific PE —> Global PH (changes marked in greyscale)

```
!!! […]
MODEL:
%WITHIN%
!!! […]
TURN_W ON GPH_W (TG_W);
TURN_W ON HPE_W (TH_W);
GPH_W ON HPE_W (GH_W);
HPE_W WITH MOR_W@0 JEP_W@0 JEE_W@0 JEC_W@0 EE_W@0 DIS_W@0 PD_W@0;
GPH_W ON COMP_W AUT_W POW_W MNG_W;
%BETWEEN%
!!! […]
TURN_B ON GPH_B (TG_B);
TURN_B ON HPE_B (TH_B);
GPH_B ON HPE_B (GH_B);
HPE_B WITH MOR_B@0 JEP_B@0 JEE_B@0 JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
GPH_B ON COMP_B AUT_B POW_B MNG_B;
```

M5. M1 + Global PE —> Specific PH (changes marked in greyscale)

```
!!! […]
MODEL:
%WITHIN%
!!! […]
TURN_W ON GPH_W (TG_W);
TURN_W ON HPE_W (TH_W);
GPH_W ON HPE_W (GH_W);
GPH_W WITH COMP_W@0 AUT_W@0 POW_W@0 MNG_W@0;
MOR_W JEP_W JEE_W JEC_W EE_W DIS_W PD_W ON HPE_W;
%BETWEEN%
!!! […]
TURN_B ON GPH_B (TG_B);
TURN_B ON HPE_B (TH_B);
GPH_B ON HPE_B (GH_B);
GPH_B WITH COMP_B@0 AUT_B@0 POW_B@0 MNG_B@0;
MOR_B JEP_B JEE_B JEC_B EE_B DIS_B PD_B ON HPE_B;
```

M6. M1 + Specific PE —> Specific PH (changes marked in greyscale)

```
!!! […]
MODEL:
%WITHIN%
!!! […]
TURN_W ON GPH_W (TG_W);
TURN_W ON HPE_W (TH_W);
GPH_W ON HPE_W (GH_W);
HPE_W WITH MOR_W@0 JEP_W@0 JEE_W@0 JEC_W@0 EE_W@0 DIS_W@0 PD_W@0;
GPH_W WITH COMP_W@0 AUT_W@0 POW_W@0 MNG_W@0;
MOR_W JEP_W JEE_W JEC_W EE_W DIS_W PD_W ON COMP_W AUT_W POW_W MNG_W;
%BETWEEN%
!!! […]
TURN_B ON GPH_B (TG_B);
TURN_B ON HPE_B (TH_B);
GPH_B ON HPE_B (GH_B);
HPE_B WITH MOR_B@0 JEP_B@0 JEE_B@0 JEC_B@0 EE_B@0 DIS_B@0 PD_B@0;
GPH_B WITH COMP_B@0 AUT_B@0 POW_B@0 MNG_B@0;
MOR_B JEP_B JEE_B JEC_B EE_B DIS_B PD_B ON COMP_B AUT_B POW_B MNG_B;
```

Section 11

Calculation of Contextual Effects, Standardized Effects, Effect Sizes, Indirect Effects, and Total Effects.

All calculations of contextual effects, standardized effects, effect sizes, indirect effects, and total effects are done in the MODEL CONSTRAINT section, using the labels (for predictive paths and variance) included in the syntax associated with M1 presented at the beginning of the previous section (Section 10) of the online supplements. To program these calculation, you would also need the estimates of the L1 and L2 variance of the various latent factors taken from the final retained DL-MLCFA (with isomorphism).

Contextual Effects

Contextual effects are calculated as the difference between the L2 effect and its L1 counterpart. Using the previous labels, these can be calculated in the MODEL CONSTRAINT section:

```
MODEL CONSTRAINT:
!!!! Calculation of contextual effects.
NEW (TG_C TH_C GH_C);
TG_C = TG_B - TG_W;
TH_C = TH_B - TH_W;
GH_C = GH_B - GH_W;
!!! Where the contextual effects (_C) is calculated as the L2 effect (-B) minus the L1 effect (_W)
```

Indirect Effects and Total Effects

Indirect effects are calculated as the product of the two effects that form it (predictor-mediator, and mediator-outcome), whereas total effects are the sum of the indirect and direct effect linking a predictor and an outcome. These are calculated separately for each level in the MODEL CONSTRAINT section. Importantly, indirect and total effects occurring at level 2 and involving contextual variables should be done using the previously calculated contextual effects. Below, we show the calculation for raw Level 2 effects (suitable for a climate variable) and for contextual level 2 effects (suitable for contextual variables).

```
MODEL CONSTRAINT:
!!!! Calculation of indirect and total effects based on L1, L2 and contextual effects.
NEW (IND1 IND2 IND3 TOT1 TOT2 TOT3);
!!! Level 1
IND1 = TG_W * GH_W;
TOT1 = IND1 + TH_W;
!!! Raw level 2 (climate)
IND2 = TG_B * GH_B;
TOT2 = IND2 + TH_B;
!!! Contextual level 2 (context).
IND3 = TG_C * GH_C;
TOT3  = IND3 + TH_C;
```

Standardized Effects and Effect Sizes.

Proper estimates of standardized effects need to be standardized (β) in relation to the total (L1 and L2) variance, whereas effect size (ES) indicators need to be defined based on the L1 variance of the outcome (Marsh et al. 2009, 2012; Morin et al., 2014).

Standardized effects are calculated as:
$\beta = b * SD_{predictor})/SD_{outcome}$

Effect sizes are calculated as:
$ES = b * SD_{predictor})/SD_{outcomeL1}$.

In these formulas, *b* is the level-specific unstandardized regression coefficient (which has been labelled in the previous section), $SD_{predictor}$ is the level-specific standard deviation of the predictor, $SD_{outcome}$ is the combined L1 and L2 standard deviation of the outcome, and $SD_{outcomeL1}$ is the L1 standard deviation of the outcome. All variance estimates (SD = square root of the variance) can be taken from the final DL-MLCFA solution (isomorphism).

These are calculated separately for each level in the MODEL CONSTRAINT section. Importantly, the calculation of standardized effects and effects sizes occurring at level 2 and involving contextual variables should be done using the previously calculated contextual effects. Below, we show the calculation for raw level 2 effects (suitable for a climate variable) and for contextual level 2 effects (suitable for contextual variables).

```
MODEL CONSTRAINT:
!!! Calculation of L1 standardized effects and effect sizes.
!!! Variance estimates taken from the multilevel CFA model.
NEW (S_PEPH S_PETI S_PHTI ES_PEPH ES_PETI ES_PHTI);
S_PEPH = GH_W  *(sqrt(.071)/sqrt(.526));
ES_PEPH = GH_W *(sqrt(.071) /sqrt(.492));
S_PETI = TH_W *(sqrt(.071)/sqrt(.452));
ES_PETI = TH_W *(sqrt(.071) /sqrt(.441));
S_PHTI = TG_W *(sqrt(.517)/sqrt(.452));
ES_PHTI = TG_W *(sqrt(.517) /sqrt(.441));
!!! Calculation of L2 standardized effects and effect sizes.
NEW (S_PEPH_B S_PETI_B S_PHTI_B ES_PEPH_B ES_PETI_B ES_PHTI_B);
S_PEPH_B = GH_B *(sqrt(.005 )/sqrt(.526));
ES_PEPH_B = GH_B *(sqrt(.005 ) /sqrt(.492));
S_PETI_B = TH_B *(sqrt(.005 )/sqrt(.452));
ES_PETI_B = TH_B  *(sqrt(.005 ) /sqrt(.441));
S_PHTI_B = TG_B *(sqrt(.036)/sqrt(.452));
ES_PHTI_B = TG_B *(sqrt(.036) /sqrt(.441));
!!! Calculation of contextual standardized effects and effect sizes.
NEW (S_PEPH_C S_PETI_C S_PHTI_C ES_PEPH_C ES_PETI_C ES_PHTI_C);
S_PEPH_C = GH_C *(sqrt(.005 )/sqrt(.526));
ES_PEPH_C = GH_C *(sqrt(.005 ) /sqrt(.492));
S_PETI_C = TH_C *(sqrt(.005 )/sqrt(.452));
ES_PETI_C = TH_C *(sqrt(.005 ) /sqrt(.441));
S_PHTI_C = TG_C *(sqrt(.036)/sqrt(.452));
ES_PHTI_C = TG_C *(sqrt(.036) /sqrt(.441));
```

Results

For all of these coefficients, the results will appear in the following output section:

MODEL RESULTS

	Estimate	S.E.	Two-Tailed Est./S.E.	P-Value
Within Level				
!!! […]				
Between Level				
!!! […]				
New/Additional Parameters				
TG_C	0.115	0.158	0.729	0.466
TH_C	0.356	0.482	0.737	0.461
GH_C	0.343	0.203	1.691	0.091
IND1	-1.487	0.124	-11.951	0.000
IND2	-1.380	0.441	-3.131	0.002
IND3	0.040	0.064	0.617	0.537
TOT1	-1.725	0.131	-13.132	0.000
TOT2	-1.263	0.193	-6.533	0.000
TOT3	0.395	0.428	0.924	0.355
S_PEPH	0.853	0.026	32.332	0.000
S_PETI	-0.091	0.032	-2.832	0.005
S_PHTI	-0.373	0.024	-15.501	0.000
ES_PEPH	0.882	0.027	32.332	0.000
ES_PETI	-0.092	0.032	-2.832	0.005
ES_PHTI	-0.378	0.024	-15.501	0.000
S_PEPH_B	0.255	0.037	6.859	0.000
S_PETI_B	0.012	0.047	0.250	0.802
S_PHTI_B	-0.074	0.023	-3.234	0.001
ES_PEPH_	0.263	0.038	6.859	0.000
ES_PETI_	0.012	0.047	0.250	0.802
ES_PHTI_	-0.075	0.023	-3.234	0.001
S_PEPH_C	0.032	0.018	1.725	0.085
S_PETI_C	0.035	0.048	0.737	0.461
S_PHTI_C	0.017	0.023	0.729	0.466
ES_PEPH_	0.033	0.019	1.725	0.085
ES_PETI_	0.036	0.049	0.737	0.461
ES_PHTI_	0.017	0.024	0.729	0.466

For all of these coefficients, the confidence intervals will appear in the output section:

CONFIDENCE INTERVALS OF MODEL RESULTS

	Lower .5%	Lower 2.5%	Lower 5%	Estimate	Upper 5%	Upper 2.5%	Upper .5%
Within Level							
!!! […]							
Between Level							
!!! […]							
New/Additional Parameters							
TG_C	-0.292	-0.194	-0.145	0.115	0.375	0.425	0.522
TH_C	-0.887	-0.590	-0.438	0.356	1.149	1.301	1.598
GH_C	-0.180	-0.055	0.009	0.343	0.677	0.741	0.866
IND1	-1.807	-1.731	-1.692	-1.487	-1.282	-1.243	-1.167
IND2	-2.516	-2.244	-2.105	-1.380	-0.655	-0.516	-0.245
IND3	-0.125	-0.086	-0.066	0.040	0.145	0.165	0.204
TOT1	-2.064	-1.983	-1.941	-1.725	-1.509	-1.468	-1.387
TOT2	-1.761	-1.642	-1.581	-1.263	-0.945	-0.884	-0.765
TOT3	-0.706	-0.443	-0.308	0.395	1.099	1.233	1.497
S_PEPH	0.785	0.802	0.810	0.853	0.897	0.905	0.921
S_PETI	-0.173	-0.154	-0.144	-0.091	-0.038	-0.028	-0.008
S_PHTI	-0.435	-0.421	-0.413	-0.373	-0.334	-0.326	-0.311
ES_PEPH	0.812	0.829	0.837	0.882	0.927	0.936	0.953
ES_PETI	-0.175	-0.156	-0.145	-0.092	-0.039	-0.028	-0.008
ES_PHTI	-0.441	-0.426	-0.418	-0.378	-0.338	-0.330	-0.315
S_PEPH_B	0.159	0.182	0.194	0.255	0.316	0.327	0.350
S_PETI_B	-0.108	-0.080	-0.065	0.012	0.088	0.103	0.132
S_PHTI_B	-0.134	-0.119	-0.112	-0.074	-0.037	-0.029	-0.015
ES_PEPH_	0.164	0.188	0.200	0.263	0.326	0.339	0.362
ES_PETI_	-0.110	-0.081	-0.066	0.012	0.090	0.104	0.133
ES_PHTI_	-0.135	-0.121	-0.114	-0.075	-0.037	-0.030	-0.015
S_PEPH_C	-0.016	-0.004	0.001	0.032	0.062	0.068	0.079
S_PETI_C	-0.088	-0.059	-0.044	0.035	0.114	0.130	0.159
S_PHTI_C	-0.043	-0.029	-0.022	0.017	0.056	0.063	0.078
ES_PEPH_	-0.016	-0.004	0.002	0.033	0.064	0.070	0.082
ES_PETI_	-0.089	-0.060	-0.044	0.036	0.116	0.131	0.161
ES_PHTI_	-0.044	-0.029	-0.022	0.017	0.056	0.064	0.079