,

Carnegie Mellon University

Department of Statistics & Data Science

Advanced Data Analysis Project Report

Stress and Cold Susceptibility: Comparing Z-Score and Latent Measurement Models

Erin Franke

December 10, 2025

Committee:
Weijing Tang, Advisor
Phoebe Lam, Carnegie Mellon University Department of Psychology

Abstract

Higher levels of composite stress and distressing emotional states have been linked to greater odds of developing a clinical cold following a viral inoculation (Cohen et al., 1991). However, these findings were based on a modest, racially homogeneous sample of White adults in the United Kingdom and relied on self-reported measures, potentially affected by measurement non-invariance—systematic differences in how individuals of different ages, sexes, or education levels respond to items. To address these limitations, we pool data from five viral challenge studies to create a larger and comparatively more diverse sample (N = 1,415; 77.5% White, 19.3% Black, 3.2% other races). Using the pooled data, we assess the effect of subjective stress, measured as perceived stress and negative affect, on vulnerability to the common cold. Although perceived stress is consistently measured, the negative affect items administered vary by study, with only partial overlap. Prior work has typically addressed this through z-scoring, which entails averaging each individual's available item scores, standardizing these averages within each study, and then aggregating the standardized values across studies—an approach that overlooks measurement invariance. Instead, we implement moderated nonlinear factor analysis (MNLFA) (Bauer and Hussong, 2009) to estimate latent perceived stress and negative affect scores that accommodate missing item blocks and ensure invariance across key demographic characteristics. Using these invariant scores, we find no significant associations between either perceived stress or negative affect and cold vulnerability following viral exposure.

Contents

1 Introduction **4**

2 Data **5**
2.1 Introduction to the five studies 5
2.2 Challenge 1: Partially overlapped features 7
2.3 Challenge 2: Differential item functioning 8

3 Methods **9**
3.1 Z-scoring approach 9
3.2 MNLFA 10
3.2.1 Identifying items that exhibit DIF 13
3.2.2 Treating ordinal variables as continuous 14
3.2.3 Extracting the factor scores 14
3.3 Using scores for downstream analysis 15

4 Results **17**
4.1 DIF detected in MNLFA models 17
4.2 Subjective stress and cold development 19

5 Discussion **20**

A Logistic regression results **25**

B Additional Study Information **26**

1 Introduction

Decades of longitudinal research have prospectively linked severe or chronic stressors—such as unemployment, caregiving burden, and exposure to violence—to increased risk for physical diseases, including cardiovascular diseases (Steptoe and Kivimäki, 2013), respiratory diseases (Graham et al., 1986), and some cancers (Chida et al., 2008). At least part of this association is believed to operate through perceived stress, meaning one's perception that environmental demands exceed the ability to control or manage them, and through negative affect, reflected in distressing emotional states such as anger, anxiety, or sadness. Higher levels of perceived stress and negative affect have been associated with biological indicators of disease risk, including elevated blood pressure, higher BMI, and metabolic dysregulation (Spruill, 2010). However, these correlational studies cannot distinguish whether stress and negative affect contribute to biological vulnerability to clinical disease or merely reflect greater pathogen exposure or disease risk factors. This limitation was addressed by using the viral challenge paradigm, in which healthy participants are experimentally inoculated under quarantine to hold exposure constant and test whether psychosocial factors predict who becomes ill.

Among the earliest applications of the viral challenge paradigm was the British Cold Study, which found that individuals reporting higher levels of a composite of stress and negative affect are more likely to develop clinical colds following viral exposure (Cohen et al., 1991). The associations remained significant even after accounting for demographic factors, seasonal factors, and pre-existing antibody levels, indicating that stress is linked with biological susceptibility to disease as opposed to differences in exposure.

However, we identify two statistical limitations to this study. First, the British Cold Study has a modest sample size of 399 participants, limiting the statistical power, and includes few people of color, which restricts the generalizability of findings beyond White adults in the United Kingdom. Second, the composite measure used to obtain these results is not measurement invariant, meaning it fails to account for differences in how individuals of different ages, sexes, or education levels respond to questionnaire items. As a result, associations involving the stress composite may be misleading. Observed effects may reflect differences in how items function or are interpreted across groups rather than true differences in underlying stress.

To address these limitations, we combine the British Cold Study with four additional viral challenge datasets (Pittsburgh Cold Studies 1–3 and the Pittsburgh Mind-Body Center Study), creating a larger and comparatively

more diverse sample of adults (N = 1,415; 77.5% White, 19.3% Black, 3.2% other races). Using the pooled data, we examine the effect of subjective stress, measured as perceived stress and negative affect, on vulnerability to the common cold. While perceived stress is measured consistently across studies, the item sets used to measure negative affect are unique to each study (with partial overlap between studies), creating structural missingness in the data. Psychological research commonly addresses this issue through z-scoring, which entails averaging each individual's available item scores, standardizing these averages within each study, and then aggregating the standardized values across studies. Although this approach is simple, the resulting constructs are not measurement invariant. As a result, our analysis uses moderated nonlinear factor analysis (MNLFA) (Bauer and Hussong, 2009), a latent variable approach, to obtain latent composite scores for perceived stress and negative affect. This method ensures measurement invariance across key demographic characteristics and can also accommodate the systematic missingness produced by the differing item sets used to measure negative affect. We then use logistic regression to assess the associations between the perceived stress and negative affect latent traits and cold development, controlling for relevant covariates and incorporating plausible values to account for measurement error in the latent traits. Our results do not indicate that higher perceived stress (OR = 1.11, 95% CI[0.98, 1.25]) or negative affect (OR = 1.13, 95% CI[0.99, 1.28]) increases vulnerability to the common cold. Given the widespread use of z-scoring within psychological research, we compare our results using MNLFA to those obtained using this standardization approach.

2 Data

2.1 Introduction to the five studies

Our data is aggregated across five studies conducted between 1986 and 2011, collectively known as the Common Cold Project (Laboratory for the Study of Stress, Immunity, and Disease, 2016). Each study enrolled volunteers determined to be in good health through medical examination. After completing baseline psychosocial questionnaires and biological assessments, volunteers consented to having a common cold virus injected through nasal drops. They were then quarantined for either five or six days (dependent on viral strain) and monitored for objective signs of illness. The first study, the British Cold Study (BCS), enrolled residents of Great Britain and took place from 1986 to 1989. The additional four studies recruited volunteers

from Pittsburgh, Pennsylvania between 1993 and 2011. These studies are titled Pittsburgh Cold Study 1 (PCS1), Pittsburgh Cold Study 2 (PCS2), Pittsburgh Mind-Body Center Study (PMBC), and Pittsburgh Cold Study 3 (PCS3). Table 1 provides further information on each study cohort. While the overall population across the studies remains predominantly White, more recent studies, such as PCS3, put forth additional effort to include volunteers from minority racial groups. Other nuances of the studies, including participant compensation information, viral strain, and criteria for infection and cold development, can be found in Appendix B.

Table 1: Participant demographics for each of the five studies. Aggregating the five studies allows for a larger and more diverse sample of individuals.

Statistic	**BCS**	**PCS1**	**PCS2**	**PMBC**	**PCS3**	**Overall**
Years	1986-1989	1993-1996	1997-2001	2000-2004	2007-2011	1986-2011
Participants	399	276	334	193	213	1415
% Female	61.2	54.7	52.4	50.8	42.2	53.7
Age Range (yrs)	18-54	18-55	18-54	21-55	18-55	18-55
Mean Age (yrs)	33.6	29.1	28.9	37.2	30.1	31.6
Race						
% White	100	81.1	67.7	66.8	55.9	77.5
% Black	-	15.2	29.6	27.5	37.1	19.3
% Other	-	3.7	2.7	5.7	7.0	3.2
% Develop cold	37.8	40.4	25.7	29.7	33.3	33.6

In addition to age, sex, and race, participant education level is also a part of the demographic information collected. Education level is recorded on an eight level scale in BCS and four level scale in each of the other studies, thus we reduce the BCS variable to the four level scale. Throughout our analysis, we will adjust for differences in how individuals of different age, sex, and education level respond to items. We do not include race as a moderator, as race itself isn't what would cause such differences and we do not want to encourage this false belief. Other relevant variables collected include the challenge virus that each participant was inoculated with, the season of the trial, whether or not the participant was seropositive for the challenge virus prior to the trial, the number of roommates with which the individual was housed (only applicable to BCS), and whether or not a roommate developed a cold.

We use the perceived stress scale (PSS) to measure perceived stress. The scale was originally developed in 1983, and remains a widely used

measure in psychological research for examining how different situations influence individuals' feelings and perceived stress (Cohen et al., 1983). Its ten questions ask the participant to evaluate "how you felt or thought over the last month" on a five-point scale ranging from *never* to *very often*. For example, participants are asked:

- In the last month, how often have you been upset because of something that happened unexpectedly?
- In the last month, how often have you felt nervous and stressed?
- In the last month, how often have you been able to control irritations in your life?

We flip any items with a positive connotation so that high values (e.g. very often) consistently represent higher perceived stress and low values (e.g. never) represent lower perceived stress. Missing data on the participant responses to the PSS items is minimal. Two of the 1,415 individuals did not respond to any of the ten items and thus are omitted from the reminder of our perceived stress analysis. Ten individuals skipped 1-2 items when filling out the survey; otherwise the data is complete across each of the five studies.

2.2 Challenge 1: Partially overlapped features

We measure negative affect using trait adjective data, which asks participants to indicate "how accurately the trait describes you as you typically are" on a five-point scale. Participants from PCS1 and PCS2 completed the survey 7-8 weeks pre-quarantine, while the survey was administered two weeks prior to quarantine in PCS3. The timeline of the questionnaire differs slightly in the BCS and PMBC studies. BCS participants were asked to evaluate to what degree they have felt each trait over the past week. PMBC participants were asked on each day of the trial how accurately each trait describes them, resulting in an average score across the 5-6 day period. We acknowledge that the studies implementing these surveys at different times relative to the quarantine period could be a concern, however, the underlying construct of negative affect is conceptualized as relatively stable. Therefore, these procedural variations are not expected generate systematic item-level bias across studies.

The primary challenge with the negative affect data is that the items administered vary by study, creating structural missingness in the combined dataset. The 25 emotions and their presence in the five studies are shown in Figure 1. Only 9 of the 25 emotions are administered to BCS participants,

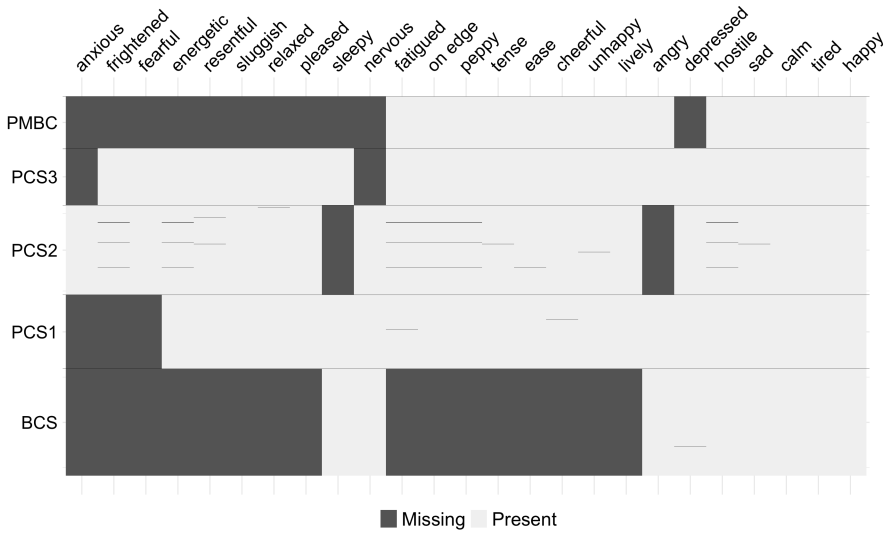

Figure 1: Structural missingness is present in the negative affect data, posing a challenge for data integration.

while 23 are administered in PCS2 and PCS3. Figure 1 also displays that a few individuals skipped responding to particular items. However, over 99% of individuals completed all items on their questionnaire. Therefore we do not consider this small amount of non-response bias a concern. In a similar manner to our procedure with the perceived stress data, we flip the values of positive traits (e.g. happy, calm, cheerful) so that higher values consistently represent an increase in negative emotion.

2.3 Challenge 2: Differential item functioning

Prior approaches to analyzing the relationship between stress and cold diagnosis have failed to obtain measurement invariant composite measures. Measurement invariance implies that the distribution of observed item responses is not affected by any variables other than the latent construct of interest (Mellenbergh, 1989). An item that violates measurement invariance is said to exhibit differential item functioning (DIF). For example, the perceived stress item asking, "In the last month, how often have you felt nervous and stressed?" will exhibit DIF if, controlling for an individual's true level of perceived stress, their probability of endorsing this item differs systematically by sex, age, education level, or another background variable. Figure 2 plots

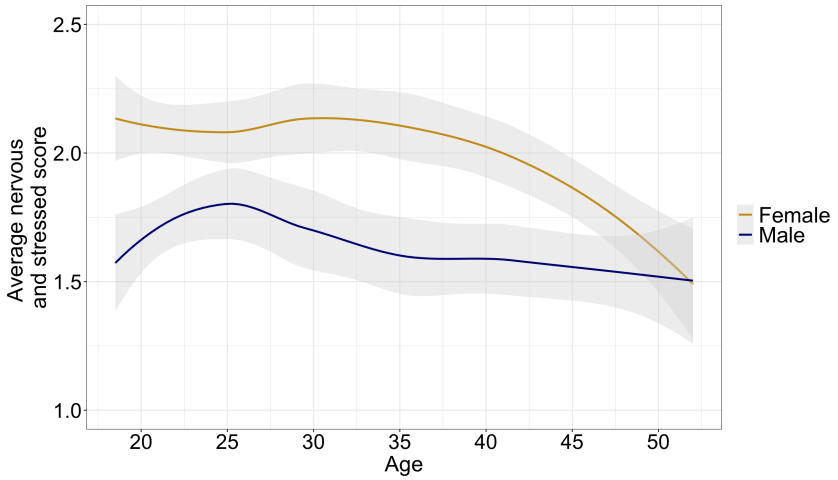

Figure 2: On average, across all ages, females report higher levels of nervousness and stress than males. However, it is unclear whether this reflects true differences in nervousness and stress or systematic differences in reporting.

the average score on the nervous and stressed item by age and sex. While it is evident that females report higher levels of nervousness and stress across all ages, it is unclear whether this reflects true differences in nervousness and stress or systematic differences in reporting. For example, it is plausible that males are less willing to admit than females that they are nervous and stressed when taking the survey. Failure to correct for differential item functioning causes bias in the composite measure, as some individuals' scores overestimate or underestimate their true level of the latent trait.

3 Methods

3.1 Z-scoring approach

In psychological research, a standard approach for forming composite measures with data from multiple studies is z-scoring. Z-scoring solves partially overlapped features by taking the average score across all items present per individual and standardizing these average scores within-study. The standardized scores are then aggregated across studies, as they are considered to be comparable measures of the latent trait.

While this method is a simple solution to structural missingness introduced through data integration, it has two flaws. First, by standardizing

scores first within-study, an assumption is made that any between-study variation in the mean or variance of the average item score is a direct result of measurement differences. In other words, the standardization step adjusts for the fact that some studies administered items that are naturally more highly endorsed, while others did not. However, the standardization process does not account for the potential of true differences in the latent trait across studies. For example, if the average perceived stress level of BCS participants was higher than that of PCS1 participants, this could not be captured through z-scores. The second limitation of z-scoring, which will motivate the remainder of this paper, is that any DIF present on the items is not corrected for. As a result, the observed item responses may be affected by, and thus correlated with, some background variable. This correlation carries over to the z-scores, making the composite measure artificially correlated with the background variable at an amount larger than the true scientific level. When we include both the composite measure and background variable in a downstream regression model, this collinearity increases the standard errors of the estimated coefficients, reducing statistical power. With the focus of our analysis being inference, stable coefficient estimates and accurate standard errors are critical. Therefore, the z-scoring approach is not feasible. However, given the frequency with which this approach is used in psychological research, we implement z-scoring to obtain composite measures for perceived stress and negative affect and compare our results against a method that addresses its concerns.

3.2 MNLFA

To overcome the limitations of the z-scoring approach, we apply moderated nonlinear factor analysis (MNLFA) (Bauer and Hussong, 2009), which provides a more flexible framework suited to the structure of our data. MNLFA generalizes traditional latent variable models to allow for simultaneous correction of DIF over multiple continuous and categorical background variables. Given that age, sex, and education level may influence how individuals endorse questionnaire items for a given level of perceived stress or negative affect, MNLFA is ideal in comparison to other methods that are only able to handle one background variable at a time (e.g. multiple-group confirmatory factor analysis). Another benefit of MNFLA is its ability to handle blockwise missingness, allowing us to leverage all 25 negative affect items. We implement our models with the `OpenMx` R package (Boker et al., 2015), which uses full-information maximum likelihood (FIML) to estimate model parameters (Arbuckle, 1996). Using the observed data, FIML calculates a

likelihood for each group of individuals with the same missingness pattern, then finds the parameter values that make these observed data most probable. Model parameters are then iteratively adjusted to find the values that maximize the total likelihood across all cases. FIML assumes that the data are missing at random, which holds in our analysis as the missingness is almost entirely structural—it is not that participants neglected responding to particular items due to embarrassment, instead these items were simply not administered on their survey.

We implement two separate MNLFA models to obtain one-dimensional latent factors for perceived stress and negative affect. In both models, the distribution of item responses is allowed to be moderated by $\mathbf{w}_i = (w_{1i}, w_{2i}, w_{3i}, w_{4i})^\top$, the vector of background variables (sex, age, age^2, education). Age is included quadratically because item response patterns may change nonlinearly across the adult age range, with different patterns of variation occurring in early versus later adulthood. The MNLFA model is structured as follows:

$$X_{ij} = b_{ij} + \lambda_{ij} f_i + e_{ij} \tag{1}$$

$$b_{ij} = b_j + \sum_{z=1}^{4} \rho_{1w,j} w_{zi} \tag{2}$$

$$\lambda_{ij} = \lambda_j + \sum_{z=1}^{4} \rho_{2w,j} w_{zi} \tag{3}$$

$$e_{ij} | f_i, \mathbf{w}_i \sim N(0, \theta_{ij}) \tag{4}$$

$$\theta_{ij} = \theta_j \exp\left(\sum_{z=1}^{4} \rho_{3w,j} w_{zi}\right) \tag{5}$$

$$f_i | \mathbf{w}_i \sim N(\mu_i, \phi_i) \tag{6}$$

$$\mu_i = \beta_0 + \sum_{z=1}^{4} \beta_w w_{zi} \quad \text{where we fix } \beta_0 = 0 \tag{7}$$

$$\phi_i = \gamma_0 \exp\left(\sum_{z=1}^{4} \gamma_w w_{zi}\right) \quad \text{where we fix } \gamma_0 = 1 \tag{8}$$

where i indexes the individual and j indexes the item. In Equation 1, X_{ij} is a $n \times p$ matrix of observed responses on each of the p items for each of the n individuals. The baseline intercept is represented by b_{ij}, λ_{ij} is the baseline loading, f_i is the normally distributed latent factor, and e_{ij} is a normally

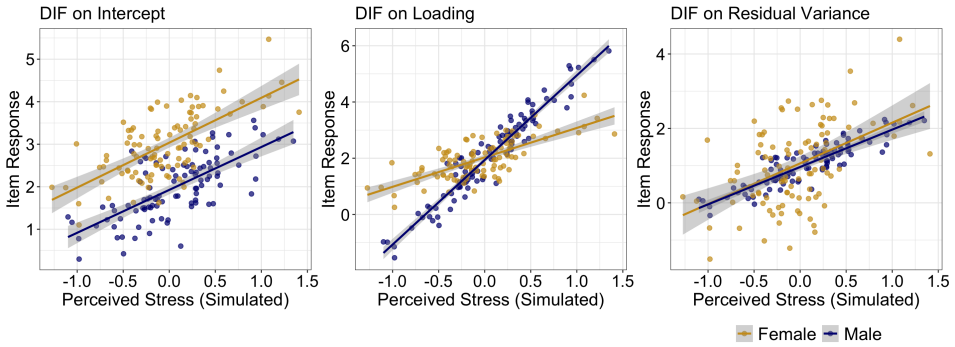

Figure 3: DIF may manifest in three ways: on the intercept, loading, and residual variance. These plots use simulated data to visualize each way that DIF may present itself.

distributed error term capturing item specific variance left unexplained by the latent factor.

DIF may present itself in three ways. In Figure 3, we simulate data to demonstrate each type of DIF, using sex as a background variable. First, DIF may be found on the intercept. Figure 3 displays that for a given level of perceived stress, females endorse this item at one unit higher than males. The MNLFA model corrects for this through Equation 2, which allows the item's baseline intercept b_j to vary with sex through $\rho_{1\text{sex},j}$. Second, Figure 3 shows how DIF may manifest in the loading. In the simulated data, a one standard deviation increase in perceived stress is associated with a three unit increase in this item in males, but only a one unit increase in females. The MNLFA model accounts for this through Equation 3, which permits the item's baseline loading λ_j to vary with sex through $\rho_{2\text{sex},j}$. By decreasing the baseline loading for females, the model now reflects that this item is a stronger indicator of perceived stress in males than it is in females. Finally, Figure 3 provides an example of what DIF on the residual variance may look like. Even after accounting for differences in perceived stress, responses to this simulated item are more variable for females than for males. Equation 4 indicates that the residuals are assumed to be normally distributed with mean zero and item-specific variance θ_{ij}. Equation 5 allows the baseline residual variance θ_j to change as a function of sex through $\rho_{3\text{sex},j}$, enabling the model to capture the greater response variability observed in females.

In addition to measurement differences across groups, there also may be structural differences in the latent trait. For example, it is possible that, on

average, perceived stress is higher and more variable in females than it is in males, after correcting for DIF. As shown in Equation 6, the latent trait is assumed to follow a normal distribution with mean μ_i and variance ϕ_i. We allow for structural differences by modeling the factor mean as a function of the background variables in Equation 7, where the coefficients β_w capture group differences in average levels of the latent trait. Similarly, Equation 8 permits the variance of the latent trait to differ across groups through coefficients γ_w. To identify the scale of the latent trait, we fix the baseline mean $\beta_0 = 0$ and the baseline variance $\gamma_0 = 1$.

3.2.1 Identifying items that exhibit DIF

To determine which items exhibit DIF, we begin by evaluating the measurement invariance of the residual variance. To do so, we first fit a model assuming strict invariance, which implies equal factor structures, factor loadings, item intercepts, and residual variances across background variables $\mathbf{w}_i$. We then conduct a series of likelihood ratio tests (LRTs), comparing the model assuming strict invariance to a model in which the residual variance for a single item is allowed to vary as a function of the four background variables. Items for which the LRT indicates a significant improvement in model fit at the Benjamini-Hochberg corrected $\alpha = 0.05$ significance level are subsequently modeled with residual variance moderation. This includes three items measuring perceived stress and twelve items measuring negative affect.

Next, we select anchor indicators, meaning items that we hold invariant across background variables. These items serve to identify the model's scale across groups defined by the background variables and to separate structural differences from measurement differences. We apply the rank-based strategy (Woods, 2008) by selecting the items that show the weakest evidence of DIF as anchor indicators. The rank-based strategy recommends suggesting roughly 20% of items to serve as anchor indicators, so we select two items for the perceived stress model and five items for the negative affect model. To identify which items show the weakest evidence of DIF, we first fit a model assuming scalar-and-metric invariance, which allows residual variance moderation but implies equal factor structures, factor loadings, and item intercepts. Using a series of LRTs, we compare this model assuming scalar-and-metric invariance to a model in which the intercept and loading for a single item are allowed to vary as a function of the four background variables. The items that yield LRTs with the largest p-values are selected as anchor indicators. We acknowledge that if these anchor indicators do exhibit DIF,

we may obtain biased parameter estimates and overestimate the amount of DIF (Wang, 2004).

Finally, we identify which items exhibit DIF on the intercept or loading. To do so, we run another series of LRTs, comparing a model that allows intercept and loading moderation on all items that are not anchors to a more constrained model where an additional item is treated as scalar-and-metric invariant. A significant LRT at the Benjamini-Hochberg corrected $\alpha = 0.05$ significance level indicates that the item is not scalar-and-metric invariant, and thus we allow the background variables to moderate the item's intercept and loading in the final MNLFA model.

3.2.2 Treating ordinal variables as continuous

All items used to measure perceived stress and negative affect are recorded on a five-level Likert scale. However, our MNLFA model treats these items as continuous by assuming each observed category reflects an underlying continuous response following a multivariate normal distribution. We choose to treat these items as continuous because the computation time for ordinal MNLFA models increases substantially. Additionally, the optimization of the likelihood function is numerically more challenging by treating these items are ordinal (Kolbe et al., 2024). While this multivariate normal distribution is misspecified with ordinal items, by consistency of the MLE, parameter estimates converge to the best fitting multivariate normal model for the ordinal data (White, 1982; Robitzsch, 2020). Furthermore, Rhemtulla et al. (2012) identified through simulation studies that treating items with at least five categories is defensible, unless in the case of asymmetric category thresholds or when estimating large factor loadings. Neither of these cases are a major concern in our data.

3.2.3 Extracting the factor scores

By Equation 9, given the background variables, we assume that latent factor and observed item responses follow a joint multivariate normal distribution.

$$\begin{pmatrix} \mathbf{x}_i \\ f_i \end{pmatrix} \Big| \mathbf{w}_i \sim N\left(\begin{bmatrix} \boldsymbol{b}_i + \boldsymbol{\Lambda}_i \mu_i \\ \mu_i \end{bmatrix}, \begin{bmatrix} \phi_i \boldsymbol{\Lambda}_i \boldsymbol{\Lambda}_i^\top + \boldsymbol{\Theta}_i & \phi_i \boldsymbol{\Lambda}_i \\ \phi_i \boldsymbol{\Lambda}_i^\top & \phi_i \end{bmatrix} \right) \tag{9}$$

We extract the factor scores according the Expected a Posteriori (EAP) scoring method (Thissen et al., 2001; Bauer and Hussong, 2009). The EAP score is defined as the mean of the posterior distribution of the latent trait for individual i given i's vector of item responses, and can be calculated using

Equation 10. The EAP is a "shrunken" estimate, meaning that the score obtained for person i will be closer to the marginal mean of the distribution of the latent trait across all individuals as the number of observed items for person i decreases (Bauer and Hussong, 2009). Applied to our data, this means we expect individuals in the British Cold Study (who only had nine negative affect items included on their survey) to have negative affect factor scores closer to the mean, on average. We are also more uncertain about these estimated scores, and therefore expect their posterior variance, given by Equation 11, to be larger.

$$\mathbb{E}[f_i|\mathbf{x}_i, \mathbf{w}_i] = \mu_i + \phi_i \mathbf{\Lambda}_i^\top (\phi_i \mathbf{\Lambda}_i \mathbf{\Lambda}_i^\top + \mathbf{\Theta}_i)^{-1} (\mathbf{x}_i - \boldsymbol{b}_i + \mathbf{\Lambda}_i \mu_i) \tag{10}$$

$$\mathbb{V}[f_i|\mathbf{x}_i, \mathbf{w}_i] = \phi_i - \phi_i^2 \mathbf{\Lambda}_i^\top (\phi_i \mathbf{\Lambda}_i \mathbf{\Lambda}_i^\top + \mathbf{\Theta}_i)^{-1} \mathbf{\Lambda}_i \tag{11}$$

3.3 Using scores for downstream analysis

We use the factor scores obtained through MNFLA for perceived stress and negative affect to fit a logistic regression model to predict cold diagnosis. Two logistic regression models are fit, one model with the covariate of interest being perceived stress and another with the covariate of interest being negative affect. We then repeat this process using the z-scores to compare results between methods. Across all four models, additional covariates include potential confounding variables in the relationship between stress and cold diagnosis—age, sex, and education level—as well as ponderal index, identity of the challenge virus, the season, serologic status for the experimental virus before the challenge, the number of participants housed together, and whether a participant housed in the same living space was infected.

Both the factor scores and z-scores are a noisy estimate of the true latent trait due to measurement error. A 10 or 25 item questionnaire cannot provide a precise measurement of one's negative affect or perceived stress. These latent traits are unobservable and can only be estimated. Thus, even while we have corrected for DIF using the MNLFA approach, measurement error is inevitable. With both the factor scores and z-scores, if we were to treat scores as observed in the logistic regression, the coefficient on perceived stress or negative affect would be attenuated due to the increased variability present from measurement error.

To solve this, we use plausible values (Mislevy, 1991). Plausible values were developed for work in large-scale assessment to obtain consistent estimates of population characteristics when students are not administered enough questions to allow for precise estimates of their ability. We take the

following steps to implement plausible values for our two logistic regression models that use factor scores extracted through MNLFA:

1. For each individual, draw from the posterior distribution of f_i to yield $M = 20$ sets of plausible values for the unobserved latent trait. The posterior distribution of f_i is a normal distribution with mean given by Equation 10 and variance given by Equation 11.

2. Use multiple imputation rules (Rubin, 1987) to obtain a confidence interval and p-value for the logistic regression coefficient of the latent trait. We implement Rubin's rules as follows:

   (a) For j in the $M = 20$ sets of plausible values, fit a logistic regression model to obtain $\hat{Q}_i$ and $\hat{U}_i$, the point and variance estimate, respectively, on the coefficient on the latent trait.

   (b) Calculate the within-imputation variance:

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m}\hat{U}_i \tag{12}$$

   (c) Calculate the between-imputation variance:

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})^2 \tag{13}$$

   (d) Calculate the total variance associated with $\bar{Q}$:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B \tag{14}$$

   (e) The statistic $(Q - \bar{Q})T^{-1/2}$ is approximately distributed as a t-distribution with degrees of freedom

$$v_m = (m-1)\left[1 + \frac{\bar{U}}{(1+m^{-1})B}\right]^2 \tag{15}$$

   (f) A 95% CI for $\bar{Q}$ can be constructed as

$$\bar{Q} \pm t_{v_m,0.975}\sqrt{T} \tag{16}$$

The resulting point estimate $\bar{Q}$ is unbiased and has standard errors that reflect both measurement error in the latent trait and uncertainty across imputations. Plausible values cannot be implemented using the z-scoring approach, as there is no measure of uncertainty associated with each z-score. Often, these z-scores are treated as objective measures without measurement error in downstream analyses, which is another limitation to this approach.

We perform diagnostic checks on the two logistic regression models fit with the EAP scores, as well as on the two models fit using the z-scores. Across all models, we add a natural spline on ponderal index to correct for nonlinearity and confirm other continuous predictors have a linear relationship with the log-odds of cold development. We choose to remove two individuals with negative affect EAP scores larger than four standard deviations above the mean, as we found them to be overly influential using Cook's distance. Randomized quantile residual plots and Q-Q plots show no signs of model misspecification for any of the four models.

4 Results

4.1 DIF detected in MNLFA models

Using MNLFA, we detect DIF on two of the ten perceived stress items. These items and their corresponding estimated moderator effects are displayed in Figure 4. Similarly, we detect DIF on 7 of 25 negative affect items, which are displayed in Figure 5. For context, we provide an interpretation of a few of these moderator effects:

- For a given level of perceived stress, a one standard deviation increase in age is associated with a 0.09 decrease in the endorsement of the item *nervous and stressed*, on average.

- A one standard deviation increase in perceived stress is associated with a 0.08 larger increase in the endorsement of the item *confident in ability to handle personal problems* for an individual with one unit higher of education, on average (e.g. comparing an individual with a two-year college degree to a four-year college degree).

- For a given level of negative affect, females endorse the trait *tired* by 0.27 more than males, on average.

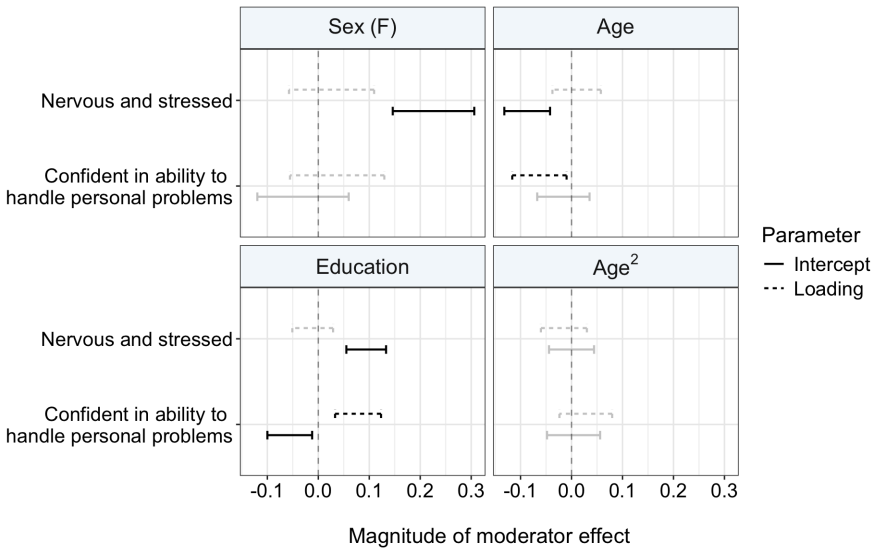

Figure 4: Estimated intercept and loading moderator effects for items exhibiting DIF for the perceived stress MNLFA model. Error bars represent the 95% CI of the effect size.

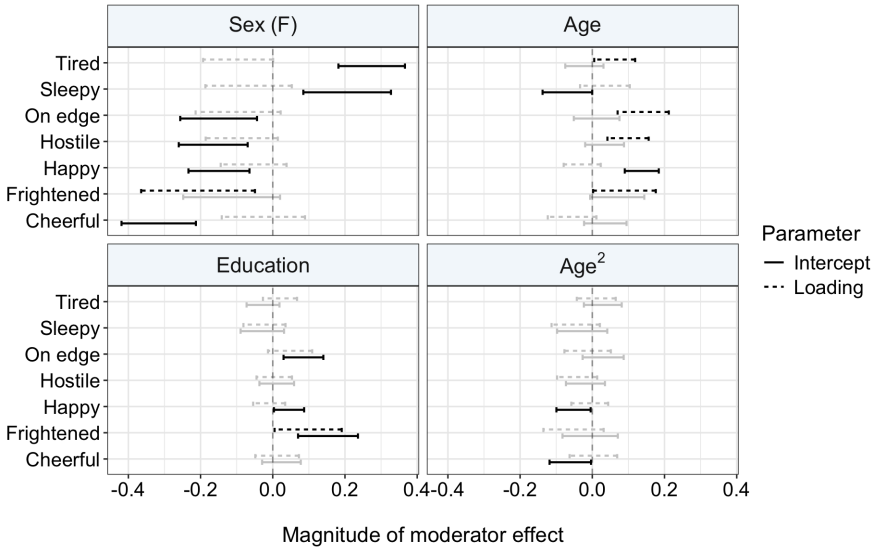

Figure 5: Estimated intercept and loading moderator effects for items exhibiting DIF for the negative affect MNLFA model. Error bars represent the 95% CI of the effect size.

4.2 Subjective stress and cold development

The results of our logistic regression models to assess the relationship between perceived stress and cold development are displayed in Table 2 (Appendix A). Using factor scores obtained through MNLFA and implementing plausible values to account for measurement error, we find that a one standard deviation increase in perceived stress causes a multiplication of the odds of developing a cold by 1.11 (95% CI [0.98, 1.25]), on average, holding all other variables constant. Using the z-scoring approach, we find that a one standard deviation increase in the perceived stress composite causes a multiplication of the odds of developing a cold by 1.06 (95% CI [0.94, 1.20]), on average.

Similarly, Table 3 (Appendix A) displays the results of our logistic regression model to understand the relationship between negative affect and cold development. Using factor scores obtained from MNLFA and implementing plausible values, we find that a one standard deviation increase in negative affect causes a multiplication of the odds of developing a cold by 1.13 (95% CI [0.99, 1.28]), on average, holding all other variables constant. With the z-scoring approach, we estimate that a one standard deviation increase in the negative affect composite causes a multiplication of the odds of developing a cold by 1.11 (95% CI [0.98, 1.25]), holding all other variables constant.

We observe that the estimated standard errors for both the perceived stress and negative affect coefficients are larger under the MNLFA approach than under the z-scoring method. A primary reason that we implemented MNLFA was because of the artificial correlation that DIF introduces into the regression model, inflating coefficient standard errors. The reason for the larger standard error observed in the MNLFA model in Table 2 and Table 3 is twofold. First, although DIF introduced correlations between the z-score and the background variables, these correlations were not strong enough to significantly inflate the coefficient standard errors in the logistic regression model. Simulations and derivations to understand how much DIF is needed to meaningful affect statistical power are still in progress, and their next steps are discussed in slightly more detail in Section 5. Second, plausible values were implemented for the latent trait derived through MNLFA, but z-scoring does not support their use. Plausible values do not treat the latent trait as an observed point estimate, which is incorrectly done in the z-scoring approach. Instead, plausible values incorporate person-specific measurement error from the MNLFA measurement model, increasing the standard error on the regression coefficient but representing a more accurate quantification of uncertainty, yielding more reliable inference.

For both z-scoring and MNLFA results, we feel comfortable drawing causal claims given that each study is designed to hold exposure constant while assessing illness susceptibility, and we account for additional confounding variables in the model. We acknowledge, however, that a one standard deviation increase in the perceived stress or negative emotion trait derived from MNLFA should reflect a one standard deviation increase in the true latent measure, while with z-scoring this cannot be assumed.

5 Discussion

Our analysis pools individuals from five viral challenge studies to understand the effect of both perceived stress and negative affect on cold vulnerability. Our work makes two key advancements on past relevant research. First, by pooling five viral challenge studies, we obtain a larger and more diverse group of individuals, increasing the generalizability of our results beyond White adults from the United Kingdom. Second, we use MNLFA to obtain measurement invariant composite measures of both negative affect and perceived stress. This ensures that our estimated effect measures the true latent trait rather then differences in how items function or are interpreted across groups. Using factor scores extracted from a MNLFA model and plausible values to account for measurement error, we do not find a significant relationship between either perceived stress and cold vulnerability (OR = 1.11, 95% CI [0.98, 1.25]) or negative affect and cold vulnerability (OR = 1.13, 95% CI [0.99, 1.28]).

One might say that our results are "borderline significant", which makes it important to consider their practical significance. We estimate that a one standard deviation in negative affect results in a multiplication of the odds of developing a cold by 1.13, holding all other covariates constant. Given that negative affect is a latent construct, this effect size can be difficult to interpret intuitively. However, an odds ratio of 1.13 does not represent a drastic change in risk of developing a cold. Even if our results were statistically significant, we do not feel that, from a public health perspective, directing resources toward reducing either perceived stress or negative affect would make a meaningful difference in the mitigation of the common cold.

While we have used data from five viral challenge studies to increase the generalizability of our results, our sample of individuals likely still does not entirely represent the general population. Given the fact that the individuals needed to take time to quarantine for several days, it is plausible that these individuals are more likely to be unemployed than the general population and

are drawn to the study for its compensation. Additionally, these individuals are known to be in good health and likely do not have anxiety surrounding catching common cold viruses, which cannot be said about the population as a whole. Furthermore, our results should not be generalized to diseases beyond the common cold, such as cancer or cardiovascular diseases. This type of paradigm that holds exposure fixed would not be ethical given the severity of those diseases. Future work on such diseases will have to take additional considerations into how to account for confounding factors and how to assess perceived stress and negative affect over many years, as these diseases take time to develop.

Our analysis also has statistical limitations. While we have corrected for DIF across age, sex, and education level, we have no way of knowing if other types of DIF are present. DIF left uncorrected may bias our extracted factor scores and our downstream results. Additionally, while past research has justified treating our five-level items as continuous, we acknowledge that a discrete factor model may be more suitable.

It is important to reiterate that our results are for perceived stress and negative affect. Other types of stress, such as objective stress (e.g. a divorce, a death in the family) may have a significant effect on vulnerability to the common cold. Furthermore, this analysis does not measure objective biological stress (e.g. cortisol levels). In health psychology, standard practice is to view stress as a construct and stress's impact on intermediate biological mechanisms as a *result* of stress, but not "biological stress". In fact, there's a wide person-to-person variation in cortisol stress reactivity for a given level of perceived stress or negative affect, supporting that understanding biological mechanisms should be considered in an entirely separate analysis.

While the primary goal of this report is to understand the effect of perceived stress and negative affect on cold vulnerability, an underlying narrative is the comparison of z-scoring and MNLFA approaches. As discussed in Section 3.1, when implementing z-scoring, the presence of DIF can reduce statistical power. There is also no straightforward way to reflect the uncertainty present in z-scores in a downstream regression. Future work will take further steps to understand the differences between MNLFA and z-scoring through simulations and derivations. We specifically plan to assess how the power of the two methods compares for differing sample sizes, levels of structural missingness, number of items with DIF present, and amount of DIF present per item. Alongside these results, we plan to provide open access to several measurement invariant psychological constructs using additional items administered on participant questionnaires during the studies. Examples of these additional constructs include social support, relationship

quality, self-worth, and agreeableness. By releasing these constructs alongside our methodological results, our hope is that researchers using the Common Cold Project data have easy access to measurement invariant scores and also understand the statistical limitations of z-scoring approaches for integrative data analysis.

References

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data.

Bauer, D. J. and Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, 14(2):101–125.

Boker, S. M., Neale, M. C., Maes, H. H., Spiegel, M., Brick, T. R., Estabrook, R., Bates, T. C., Gore, R. J., Hunter, M. D., Pritikin, J. N., Zahery, M., and Kirkpatrick, R. M. (2015). Openmx: Extended structural equation modelling.

Chida, Y., Hamer, M., Wardle, J., and Steptoe, A. (2008). Do stress-related psychosocial factors contribute to cancer incidence and survival? *Nature Clinical Practice Oncology*, 5(8):466–475.

Cohen, S., Kamarck, T., and Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*, 24(4):385.

Cohen, S., Tyrrell, D. A., and Smith, A. P. (1991). Psychological stress and susceptibility to the common cold. *New England Journal of Medicine*, 325(9):606–612.

Graham, N. M. H., Douglas, R. M., and Ryan, P. (1986). Stress and acute respiratory infection. *American Journal of Epidemiology*, 124(3):389–401.

Kolbe, L., Molenaar, D., Jak, S., and Jorgensen, T. D. (2024). Assessing measurement invariance with moderated nonlinear factor analysis using the r package openmx. *Psychological Methods*, 29(2):388–406.

Laboratory for the Study of Stress, Immunity, and Disease (2016). Common cold project. Retrieved from http://www.commoncoldproject.com.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, 13(2):127–143.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2):177–196.

Rhemtulla, M., Brosseau-Liard, P. , and Savalei, V. (2012). When can categorical variables be treated as continuous? a comparison of robust continuous and categorical sem estimation methods under suboptimal conditions. *Psychological Methods*, 17(3):354–373.

Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education*, 5.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Spruill, T. M. (2010). Chronic psychosocial stress and hypertension. *Current Hypertension Reports*, 12(1):10–16.

Steptoe, A. and Kivimäki, M. (2013). Stress and cardiovascular disease: An update on current knowledge. *Annual Review of Public Health*, 34(1):337–354.

Thissen, D., Nelson, L., Rosa, K., and McLeod, L. D. (2001). *Item response theory for items scored in two categories*. Routledge.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of rasch models. *The Journal of Experimental Education*, 72(3):221–261.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1.

Woods, C. M. (2008). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, 33(1):42–57.

A Logistic regression results

Table 2: Using both MNLFA factor scores and z-scores to measure perceived stress, we do not find that higher perceived stress results in an increased likelihood of cold development. The reported odds ratio and standard error for perceived stress under MNLFA are calculated using plausible values.

	MNLFA			**Z-Scoring**		
Covariate	OR	SE	p-value	OR	SE	p-value
(Intercept)	2.12	0.752	0.317	2.15	0.75	0.309
Perceived stress score	**1.11**	**0.064**	**0.116**	1.06	0.061	0.329
Age	1.02	0.006	0.012*	1.02	0.006	0.012*
Female	1.05	0.127	0.685	1.06	0.127	0.653
Ponderal index (spline 1)	0.39	0.436	0.031*	0.39	0.436	0.030*
Ponderal index (spline 2)	1.46	0.660	0.5645	1.46	0.660	0.566
Ponderal index (spline 3)	0.66	1.182	0.723	0.65	1.182	0.712
Ponderal index (spline 4)	1.37	1.274	0.803	1.32	1.276	0.827
Education *(ref: HS grad or lower)*						
Some college, but < 2 yrs	0.94	0.170	0.697	0.94	0.170	0.697
2+ yrs college and degree	0.70	0.207	0.084	0.70	0.207	0.078
Bachelor's degree or higher	0.84	0.171	0.310	0.84	0.171	0.311
Challenge virus *(ref: coronavirus type 229E)*						
Rhinovirus type 21	0.68	0.577	0.505	0.67	0.576	0.482
RSV	0.28	0.483	0.008**	0.28	0.483	0.008**
Rhinovirus type 14	0.26	0.398	< 0.001***	0.26	0.398	0.001***
Rhinovirus type 2	0.15	0.414	< 0.001***	0.15	0.414	< 0.001***
Rhinovirus type 23	0.20	0.608	0.0073**	0.20	0.608	0.007**
Rhinovirus type 39	0.56	0.545	0.285	0.55	0.543	0.268
Rhinovirus type 9	0.28	0.373	< 0.001***	0.28	0.373	0.001***
Seropositive	0.27	0.135	< 0.001***	0.27	0.135	< 0.001***
Season *(ref: Winter)*						
Spring	0.76	0.192	0.153	0.76	0.193	0.163
Summer	0.68	0.223	0.088	0.68	0.223	0.089
Fall	0.98	0.212	0.912	0.99	0.212	0.948
Number of roommates *(ref: 0)*						
One roommate	1.49	0.546	0.465	1.48	0.546	0.475
Two roommates	0.91	0.579	0.867	0.90	0.579	0.855
Roommate infected	1.89	0.339	0.061	1.89	0.339	0.061·

Table 3: Using both MNLFA factor scores and z-scores to measure negative affect, we do not find that higher negative affect results in an increased likelihood of cold development. The reported odds ratio and standard error for negative affect under MNLFA are calculated using plausible values.

	MNLFA			**Z-Scoring**		
Covariate	OR	SE	p-value	OR	SE	p-value
(Intercept)	2.11	0.745	0.325	2.25	0.76	0.282
Negative affect score	**1.13**	**0.067**	**0.077**	1.11	0.061	0.098
Age	1.02	0.006	0.009**	1.02	0.006	0.0096**
Female	1.07	0.126	0.587	1.08	0.126	0.562
Ponderal index (spline 1)	0.39	0.436	0.031*	0.38	0.436	0.028*
Ponderal index (spline 2)	1.46	0.661	0.569	1.45	0.660	0.575
Ponderal index (spline 3)	0.63	1.185	0.693	0.62	1.184	0.682
Ponderal index (spline 4)	1.34	1.280	0.820	1.32	1.280	0.829
Education *(ref: HS grad or lower)*						
Some college, but < 2 yrs	0.93	0.171	0.655	0.93	0.170	0.659
2+ yrs college and degree	0.69	0.206	0.070	0.69	0.206	0.067
Bachelor's degree or higher	0.83	0.172	0.284	0.83	0.172	0.280
Challenge virus *(ref: coronavirus type 229E)*						
Rhinovirus type 21	0.69	0.581	0.515	0.65	0.580	0.451
RSV	0.26	0.487	0.006**	0.26	0.487	0.0053**
Rhinovirus type 14	0.26	0.403	< 0.001***	0.26	0.403	0.0007***
Rhinovirus type 2	0.14	0.418	< 0.001***	0.14	0.418	< 0.001***
Rhinovirus type 23	0.21	0.614	0.010*	0.19	0.611	0.0067**
Rhinovirus type 39	0.58	0.551	0.316	0.53	0.547	0.247
Rhinovirus type 9	0.26	0.378	< 0.001***	0.26	0.377	0.0004***
Seropositive	0.27	0.135	< 0.001***	0.27	0.135	< 0.001***
Season *(ref: Winter)*						
Spring	0.75	0.193	0.128	0.75	0.193	0.136
Summer	0.68	0.223	0.079	0.67	0.224	0.078
Fall	0.96	0.213	0.854	0.97	0.213	0.881
Number of roommates *(ref: 0)*						
One roommate	1.53	0.548	0.437	1.52	0.548	0.446
Two roommates	0.89	0.581	0.841	0.89	0.581	0.840
Roommate infected	1.92	0.340	0.055	1.92	0.340	0.055

B Additional Study Information

The British Cold Study (BCS) took volunteers to participant in trials at the Medical Research Council's Common Cold Unit (CCU) in Salisbury, England. Travel expenses were reimbursed for all participants and they were provided with meals and accommodations during the study. Participants received nasal drops containing one of five respiratory viruses, displayed in Table 4. Cold infection criteria includes either a significant increase in serum specific IgG or IgA pre- to post-viral challenge, a 4-fold increase in serum

neutralizing Ab titer to challenge virus, or any post viral shedding. Provided the participant met criteria for infection, criteria for cold development was built on physician diagnosis based on daily cold symptom protocols.

Participants in each of the Pittsburgh Cold Studies (PCS1, PCS2, and PCS3) were recruited from newspaper advertisements in the Pittsburgh, Pennsylvania metropolitan area. Participants in PCS1 and PCS2 were compensated $800, while those in PCS3 were compensated $1000, plus an additional $60 if they provided hair samples for cortisol analysis. Participants in the Pittsburgh Mind-Body Center (PMBC) Study responded to advertisements and were compensated $800. To maximize the rate of infection, in PCS3 and PMBC only eligible volunteers with viral-specific antibody titers ≤ 4 were included in the study. Table 4 displays the challenge virus administered by nasal drops to participants in each study. Participants met criteria for infection if they underwent seroconversion or viral shedding. Participants met objective criteria for cold development if they met criteria for infection and had either a total adjusted mucus weight of greater than ten grams or an average adjusted nasal clearance time of at least seven minutes.

Table 4: Number of study participants receiving a particular strain of challenge virus. Rhinovirus 39 was the most commonly administered viral strain across studies.

Virus	BCS	PCS1	PCS2	PMBC	PCS3	Total
Coronavirus type 229E	55	0	0	0	0	55
Rhinovirus type 2	86	0	0	0	0	86
Rhinovirus 9	126	0	0	0	0	126
Rhinovirus 14	92	0	0	0	0	92
Rhinovirus 39	0	147	228	193	213	781
Rhinovirus 21	0	129	0	0	0	129
Rhinovirus 23	0	0	106	0	0	106
RSV	40	0	0	0	0	40