Comprehensive Psychoneuroendocrinology 5 (2021) 100025

Contents lists available at ScienceDirect

Comprehensive Psychoneuroendocrinology

journal homepage: www.elsevier.com/locate/cpnec

Allostatic load scoring using item response theory

Shelley H. Liu [a,*], Robert-Paul Juster [b], Kristen Dams-O'Connor [c], Julie Spicer [d]

[a] Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, United States
[b] Department of Psychiatry and Addiction, University of Montreal, Canada
[c] Department of Rehabilitation Medicine, Department of Neurology, Icahn School of Medicine at Mount Sinai, United States
[d] Department of Psychiatry, Icahn School of Medicine at Mount Sinai, United States

ARTICLE INFO

Keywords:
Allostatic load
Item response theory
National health and nutrition examination survey
Biomarkers
Psychometrics
Depression

ABSTRACT

Allostatic load is commonly operationalized using a sum-score of high-risk biomarkers. However, this method implies that biomarkers contribute equally to allostatic load, as each is given equal weight. Our goal in this methodological paper is to evaluate this, and complementarily, to identify biomarkers that are most informative and least informative for developing an allostatic load index. Item response theory models provide an alternate approach to calculating the allostatic load score, by treating individual biomarkers (e.g. "items") as indicators of a latent allostatic load construct. Item response theory scores account for the data-driven discriminating power of each biomarker, and an individual's pattern of biomarker responses. To demonstrate feasibility of this approach, we used data from the 2015–2016 National Health Examination and Nutrition Survey (NHANES; $N = 3751$), with twelve allostatic load biomarkers representing immune response, metabolic function and cardiovascular health. Item response theory models revealed that body-mass-index and C-reactive protein were the most informative biomarkers for allostatic load. Both higher allostatic load sum-score and allostatic load item response theory score were associated with lower socio-economic status ($p = 0.008$; $p<0.001$, respectively). Further, both formulations of allostatic load were positively associated with a nine-item depression screener ($p<0.001$ for both), but only the item response theory score was also positively associated with the impact of depressive symptoms on daily life ($p = 0.045$). Item response theory scores may be more finely tuned to tease out effects, compared to sum-scores, and also provide more flexibility when there are missing biomarker measurements. Supplemental R code for our approach are included.

1. Introduction

Allostatic load is a latent construct defined as multi-systemic physiological dysregulation. Allostatic load can be indexed to quantify the 'wear and tear' on the body, by quantifying multiple biomarkers representing hypothalamic-pituitary-adrenal (HPA) axis, immune/inflammation, cardiovascular, lipid and glucose functioning, as well as emergent biomarkers. Central to advances in the field of psychoneuroendocrinology, allostatic load is associated with both psychosocial exposures and health outcomes including socioeconomic status [1], race/ethnicity [1], sexual orientation [2,3], workforce burnout [4,5], aging and mortality [6], perinatal outcomes [7] and psychiatric emergencies [8].

There has been much discussion regarding the measurement of allostatic load and the relative importance and weight of individual biomarkers. In particular, there is continuing debate on the best ways to operationalize and measure allostatic load, and what can be learned from the multiple biomarkers that comprise it. While allostatic load cannot be directly observed because it is a latent and unobservable trait, researchers typically estimate it by summarizing data from a number of appropriate clinical biomarkers. Allostatic load is commonly operationalized using a sum-score of high-risk biomarkers. This traditional count-based approach, such as that used in analyzing the MacArthur Studies of Successful Aging [9], relies on dichotomizing each biomarker as high/-low risk using sample-dependent cutoffs or clinical cutoffs, and then summing the number of high-risk biomarkers to arrive at an allostatic load summary score. However, this approach assumes that each biomarker included in the sum score makes equal contributions to overall allostatic load, by treating all of the biomarkers as though they are interchangeable. It is also assumes that the simple sum score is capable of measuring allostatic load with equal precision across the full range of the underlying latent construct.

In addition to the count-based approach applied in the MacArthur

* Corresponding author. 1 Gustave L Levy Pl, New York, NY, 10029, United States.
E-mail address: shelley.liu@mountsinai.org (S.H. Liu).

https://doi.org/10.1016/j.cpnec.2020.100025
Received 28 November 2020; Received in revised form 7 December 2020; Accepted 15 December 2020
2666-4976/© 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

studies and other international samples [10–13], alternative approaches have included a two-tailed 10th/90th percentile approach [14], a standard deviation cut-off approach [15], z statistic weights generated from bootstrapping [16], factor-analytic approaches [17], and the more recent "scaling" of multi-systemic dysregulations [18]. To date, there is no consensus on the most appropriate allostatic load formulation to use [4, 19].

Our goal here is to use advanced psychometric methods to evaluate alternate methods to calculating an allostatic load score and complementarily, to identify individual biomarkers that are most informative and least informative for developing an allostatic load index. Further, we determine if individual biomarkers provide information along the entire allostatic load continuum, or if they provide information for only a portion of the continuum. While there is no universal agreement on the specific biomarkers to include in an allostatic load calculation, researchers generally agree that cardiovascular, metabolic and immune biomarkers should be included. In this paper, we will demonstrate an alternate approach to calculating allostatic load scores with dichotomous biomarker data, using data from the National Health Examination and Nutrition Survey (NHANES) for illustration. Note that NHANES does not provide neuroendocrine biomarkers often used in allostatic load studies. Many studies (>20) have used NHANES to study allostatic load [20], finding associations with outcomes such as sleep disorders [21], all-cause mortality [22] and cognitive function [23], demonstrating that the biomarkers available in NHANES are a valid and accurate measure of allostatic load.

Item response theory (IRT) is a set of psychometric models used for measurement and the development of scales [24]. Most commonly used in the education testing literature, it has now been used in biomedical research in the development and scoring of patient reported outcomes [25,26]. To our knowledge, it has not yet been used in allostatic load biomarker data, but IRT shares similarities with factor-analytic approaches that have been previously used to investigate allostatic load scoring [17]. IRT is a class of psychometric models that can be used to explain the relationship between a latent, unobservable, trait (e.g., allostatic load burden), and observable characteristics or items (e.g. clinical biomarkers). Using IRT, the measurement properties of the clinical biomarkers, the set of individuals measured, and the allostatic load burden are linked together.

The allostatic load burden and the "items" (biomarkers) used to measure it are assumed to span an unobservable continuum. Thus, we can use IRT to establish each individual's position on that continuum (e.g., quantify each individual's allostatic load burden). Unlike the sum-score approach, biomarkers are differentially weighted in a data-driven manner, depending on how much information they provide to the overall allostatic load burden scale. Each biomarker has an estimated difficulty parameter (how likely a participant with a certain latent allostatic load level will score high-risk on that biomarker), and an estimated discrimination parameter (how informative scoring high-risk on a certain biomarker is with respect to gauging the participant's latent allostatic load level; analogous to factor loadings). IRT-based scores differentially weight the contributions of the individual biomarkers based on their difficulty and discrimination. If we visualize the latent allostatic load burden as a ruler, with participants averaging a score at 0 and each 1-unit represents a standard deviation, we can interpret the IRT allostatic load burden score as a z-score.

Identification of the parameters needed to calculate an IRT-based allostatic load score requires access to a large and representative calibration sample. In accordance, we here use a nationally representative sample of (n = 3751) US adults and use survey-weighted 25th or 75th percentile of each biomarker to define representative levels of high biomarker levels. In this paper, we dichotomize biomarkers into high/low risk and compare IRT methods versus sum-score methods for calculating allostatic load scores.

2. Methods

2.1. Analytic sample

We used data from the 2015–2016 National Health Examination and Nutrition Survey (NHANES), which is provided by the National Center for Health Statistics (NCHS) in the Centers of Disease Control and Prevention (CDC). NHANES is a recurring cross-sectional survey of the non-institutionalized civilian US population, who live in the 50 states and the District of Columbia, with details available elsewhere [27]. The study sample consisted of adults aged 20 years and older but less than 60 years. We excluded those with a positive urine pregnancy test, yielding a final sample size of n = 3751.

2.2. Allostatic load

To assess allostatic load, we included biomarkers that were most commonly used in calculations of allostatic load from NHANES data [20]. Twelve commonly used biomarkers [28] were included to represent *immune* response (high sensitivity C-reactive protein (CRP), white blood cell count), *metabolic function* (glycohemoglobin, serum albumin, serum creatinine, total cholesterol, high density lipoprotein (HDL), serum triglycerides, body mass index (BMI)), and *cardiovascular health* (average of three resting systolic blood pressure measurements, average of three resting diastolic blood pressure measurements, pulse rate). For each biomarker, we found a high-risk cutoff, which was defined as the survey-weighted 75th percentile for all biomarkers except HDL and serum albumin, for which the high-risk cutoff was defined as the survey-weighted 25th percentile. For each of the twelve biomarkers, an individual received a score of 1 if their biomarker level was more extreme than the high-risk cutoff, and 0 otherwise. The allostatic load sum-score was calculated by taking the sum of all twelve biomarker scores. Thus, the allostatic load sum-score can range from 0 to 12.

2.3. Statistical analysis

Data from the NHANES cycle was extracted using the "nhanesA: NHANES Data Retrieval" R package [29]. We linked demographic data, laboratory data and physical exam data using a unique survey participant identifier. Our analyses accounted for the NHANES complex survey design, in order for findings to be considered nationally representative of the US population. We accounted for sampling strata, cluster and weights using the "survey: Analysis of complex survey samples" R package [30].

We reported the survey weighted frequency for the categorical variables, median and interquartile range (IQR), which is the difference between the 75th and 25th percentiles, for continuous variables. We investigated the correlation between the allostatic load sum-score and individual biomarkers using Pearson correlation. We then fitted two-parameter logistic IRT models using the R package "ltm: Latent trait models under IRT" [31], to the twelve dichotomized biomarkers in order to estimate the difficulty and discrimination parameters of the biomarkers, and to estimate an allostatic load burden score (IRT score), using expected a priori scores. We then plotted the item characteristic curves, item information curves, and the test information curve. Lastly, we plotted the estimated allostatic load burden score, against the sum-score, in order to visualize the correlation and compare and contrast those scores.

To validate our formulations of allostatic load, we assessed associations between the sum-score and IRT score formulations of allostatic load with socio-economic status (SES), as measured by family income-to-poverty ratio, since allostatic load is conceptualized as physiological weathering due stressful circumstances such as low SES. We then assessed associations of sum-score and IRT score formulations of allostatic load with depression, per the Patient Health Questionnaire (PHQ9), a nine-item depression screener that assesses the frequency of depression symptoms in the past two weeks by self-report [32,33]. Negative

binomial regression models were used because there are excess zeroes in the PHQ9 scores. We adjusted for covariates of age, sex, race/ethnicity and family income-to-poverty ratio. We also assessed associations of allostatic load with impacts of depressive symptoms on daily life, as measured by the question, "How difficult have these problems [PHQ9] made it for you to do your work, take care of things at home, or get along with people?" We coded the response as binary to focus on those whose depressive symptoms had substantial impacts on their daily life (not difficult at all or somewhat difficult vs. very or extremely difficult). We used logistic regression adjusted for age, sex, race/ethnicity and family income-to-poverty ratio.

We provide a tutorial and reproducible code to implement IRT in the Supplementary Materials section VI.

3. Results

3.1. Clinical and socio-demographic characteristics

Table 1 contains the survey-weighted socio-demographic and clinical covariates. The median age of the sample was 40 [interquartile range (IQR): (29, 50)]. The sample contained equal men and women. Median family income to poverty ratio was 3.0 [IQR: (1.5, 5.0)]. The sample consisted of 59.9% Non-Hispanic Whites, 10.4% Mexican American, 7.4% Other Hispanic, 12.3% Non-Hispanic Black, 6.3% Non-Hispanic Asian and 3.7% Other Race/Multi-racial.

3.2. Allostatic load

Because we used the survey-weighted 75th percentile (or 25th percentile) cutoffs to define high-risk for each biomarker, the proportion of the sample belonging to the high-risk group was not always 25%, as would be expected if we did not use the survey-weighted cutoffs. The allostatic load sum-score ranged from 0 to 12, with median of 3 [IQR: (1, 4)]. Supplementary Fig. 1 depicts the distribution of the allostatic load sum-score in the sample. The sum-score was most strongly correlated with CRP (r = 0.523), BMI (r = 0.521) and glycohemoglobin (r = 0.506) and was least correlated with creatinine (r = 0.221). There was a moderate correlation between a few additional biomarkers (see Supplementary Fig. 2, which provides a correlation plot of the twelve biomarkers plus the allostatic load sum-score). Specifically, SBP and DBP had a correlation of 0.432, BMI and CRP had a correlation of 0.403, and triglycerides and HDL had a correlation of 0.387.

Table 1
Survey-weighted sociodemographic and clinical covariates. Interquartile range denotes the interval covering the 25th to 75th percentile. Weighted frequencies and summary statistics were calculated using the R "survey" package which accounted for NHANES complex survey design.

Covariate	Summary measure
Weighted frequency (%)	
Sex	
Male	49.9
Female	50.1
Race/Ethnicity	
Mexican American	10.4
Other Hispanic	7.4
Non-Hispanic White	59.9
Non-Hispanic Black	12.3
Non-Hispanic Asian	6.3
Other Race/Multi-Racial	3.7
Median (interquartile range)	
Age (years)	40 (29, 50)
Family income-to-poverty ratio	3.0 (1.5–5.0)
White blood cell count (1000 cells/uL)	7.1 (5.9–8.7)
C-reactive protein (mg/L)	1.7 (0.6–4.3)
Body mass index (kg/m^2)	28.2 (24.2–33.0)
Serum triglycerides (mg/dL)	118 (77–187)
Serum albumin (g/dL)	4.4 (4.2–4.6)
Serum creatinine (mg/dL)	0.82 (0.69–0.95)
Systolic blood pressure (mmHg)	118 (110–127)
Diastolic blood pressure (mmHg)	72 (65–78)
Pulse rate (beats per min)	72 (66–80)
High density lipoprotein (mg/dL)	51 (41–64)
Total cholesterol (mg/dL)	188 (164–216)
Glycohemoglobin (%)	5.4 (5.1–5.7)
Urinary creatinine (mg/dL)	113 (65–184)
Allostatic load sum-score	3 (1, 4)

3.3. Item response theory

We fit a two-parameter logistic model to the twelve dichotomized biomarkers, treating each biomarker as an "item". We evaluated item fit statistics and found that all items fit the 2 PL model. The item characteristic curves (ICCs) are presented in Fig. 1. The steeper the ICC, the more the biomarker is strongly related to allostatic load. The discrimination of the biomarkers varied – CRP and BMI provided the most discrimination; while serum creatinine provided the least discrimination. The ICC for serum creatinine is mostly flat, meaning that it does not provide much information at any range of allostatic load burden, and thus is not an informative biomarker for the IRT score for this sample. Fig. 2 contains the item information curves, which similarly demonstrates that BMI and CRP provide the most information about the allostatic load burden, and is most informative for slightly higher than average allostatic load burden levels. Supplementary Fig. 3 presents the test information curve, which shows that the test provides the most information for allostatic load burden that is 1 standard deviation above the average burden for the population, but provides less information about very low or very high allostatic load burden levels. The distribution of the IRT score is presented in Supplementary Fig. 4.

3.4. Relationship between allostatic load IRT (burden) score and allostatic load sum-score

Fig. 3 shows the plot between the sum-score and IRT score formulations of allostatic load. Although there is a general monotonic relationship (positive correlation) between the sum-score and the IRT score, we found that a high sum-score did not always imply a high IRT score, since the IRT score also depended on the biomarker characteristics (e.g., discrimination) and the response pattern (participants may be high-risk on different sets of biomarkers).

3.5. Association between socio-economic status and allostatic load sum-score and IRT score

To verify our formulation of allostatic load reflects the conceptualization of allostatic load as representative of physiological weathering due to stressful circumstances such as low SES, we assessed associations of sum-score and IRT score with SES (family income-to-poverty ratio) (Supplementary Table 1). SES was significantly negatively associated with both the sum-score and the IRT score ($p = 0.008$; $p<0.001$, respectively).

3.6. Associations between allostatic load sum-score and IRT score with depression

We then assessed adjusted associations of the sum-score and IRT score with a depression screener, the PHQ9 (Table 2). Both the sum-score and IRT score were significantly positively associated with PHQ9 ($p<0.001$ for both).

Lastly, we assessed adjusted associations of sum-score and IRT score with impacts of depressive symptoms on daily life (Table 3). Notably, only the IRT score was significantly positively associated with this outcome ($p = 0.045$), while the sum-score was not associated ($p = 0.20$).

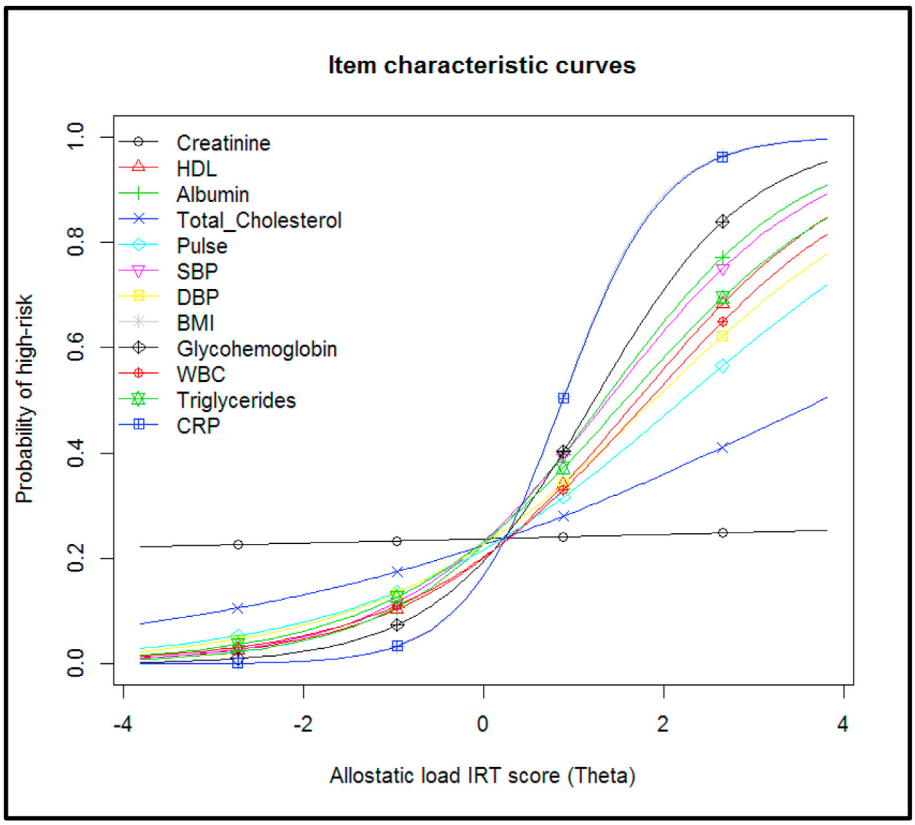

Fig. 1. Item characteristic curves for twelve allostatic load biomarkers in the NHANES 2015–2016 study, using a 2 parameter logistic model.

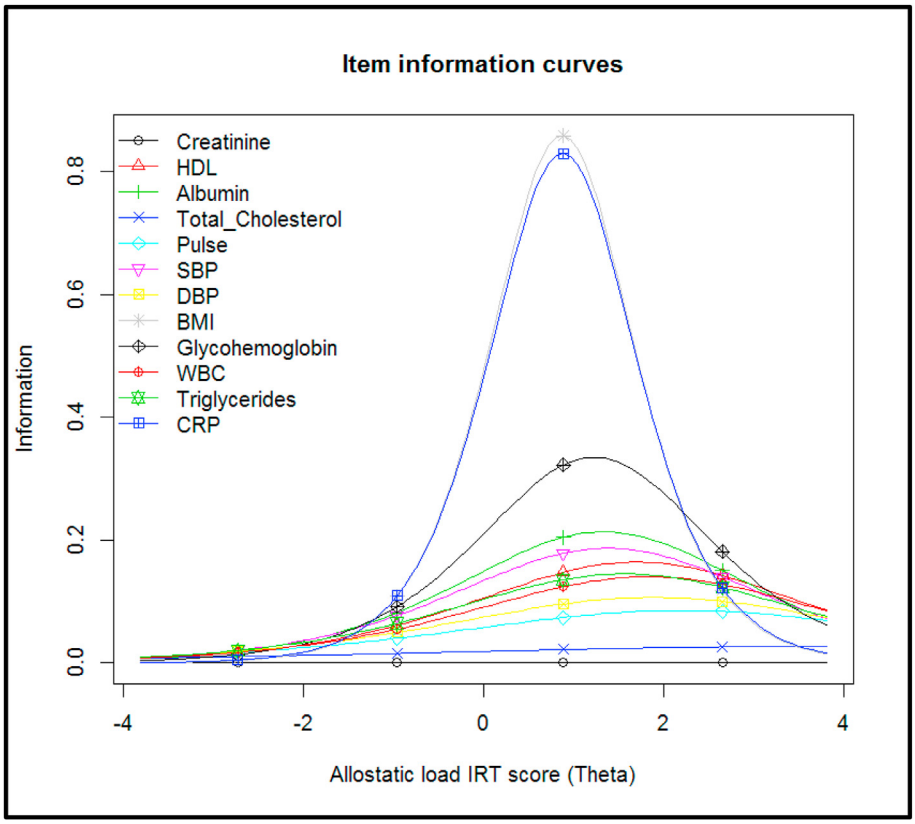

Fig. 2. Item information curves for twelve allostatic load biomarkers in the NHANES 2015–2016 study, using a 2 parameter logistic model.

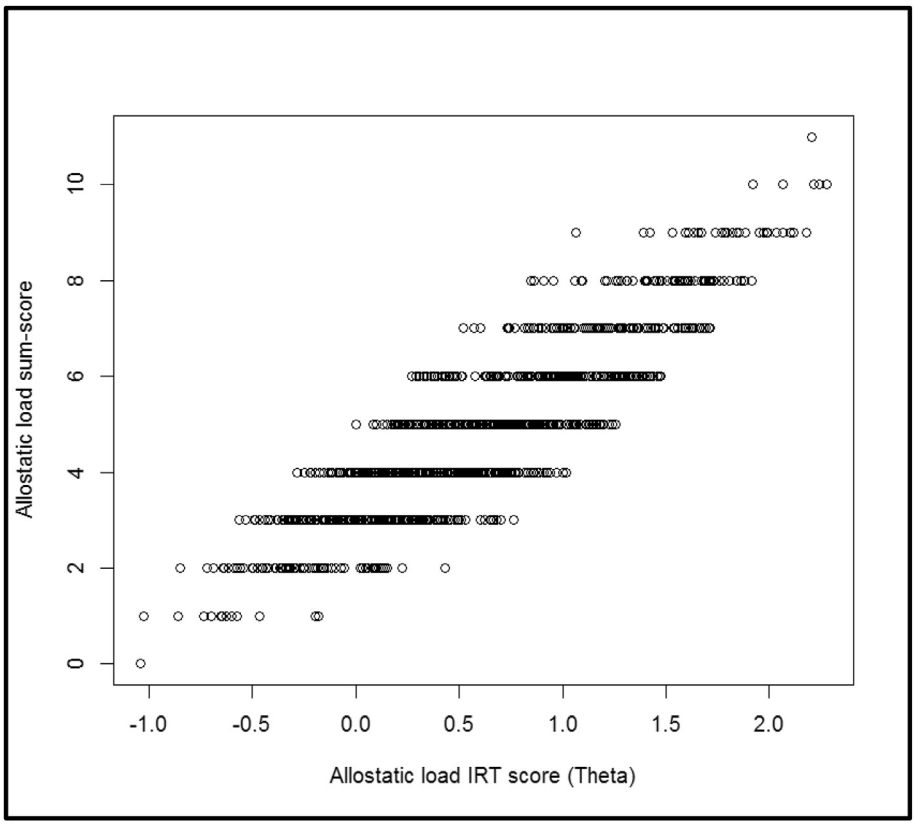

Fig. 3. Plot of the allostatic load sum-scores versus the estimated allostatic load IRT scores, for the NHANES 2015–2016 study.

Table 2
Adjusted associations of allostatic load IRT and sum-scores with depression screener (Patient Health Questionnaire, PHQ9) in the NHANES 2015–2016 study. Negative binomial regression was used to account for excess zeros in PHQ9 scores. Models were adjusted for sex, age, race/ethnicity and SES (family income-to-poverty ratio). The models have different sample sizes, as IRT does not require complete data on all allostatic load biomarkers in order to calculate scores, unlike the sum-score approach.

Predictors	*Incidence Rate Ratios*	*95% CI*	*p*	*Incidence Rate Ratios*	*95% CI*	*p*
(Intercept)	4.41	3.60 – 5.41	**<0.001**	3.69	3.00 – 4.55	**<0.001**
AL IRT Score	1.19	1.12 – 1.26	**<0.001**			
AL Sum Score				1.06	1.03 – 1.08	**<0.001**
Sex						
Male	Reference					
Female	1.27	1.15 – 1.39	**<0.001**	1.3	1.18 – 1.44	**<0.001**
Age (years)	1	1.00 – 1.01	0.493	1	1.00 – 1.01	0.618
Race/Ethnicity						
Non-Hispanic White	Reference					
Mexican American	0.69	0.60 – 0.79	**<0.001**	0.73	0.63 – 0.84	**<0.001**
Other Hispanic	0.94	0.80 – 1.09	0.409	0.99	0.84 – 1.16	0.88
Non-Hispanic Black	0.8	0.70 – 0.90	**<0.001**	0.82	0.72 – 0.95	**0.006**
Non-Hispanic Asian	0.74	0.63 – 0.86	**<0.001**	0.78	0.66 – 0.93	**0.004**
Other Race, including Multi-Racial	1.08	0.86 – 1.37	0.521	1.12	0.89 – 1.44	0.348
Family income to poverty ratio	0.86	0.84 – 0.89	**<0.001**	0.86	0.83 – 0.89	**<0.001**
Observations	2913			2581		

4. Discussion

Our study is the first to demonstrate the use of item response theory to understand how different biomarkers provide information about latent allostatic load burden. We demonstrate an alternative method to calculate allostatic load, beyond the traditional sum-score approach, which can account for discriminating abilities of individual biomarkers. Our findings suggest that BMI and CRP (or related immune measures) should be included in calculations of allostatic load. Additionally, CRP and BMI are often positively correlated [34–36], as was found in the current study. Further, our findings provide evidence that an alternative way of calculating allostatic may provide more variability in the allostatic load scores than the traditional sum-score approach. This can help tease out additional effects not seen using sum-scores if allostatic load is used as a predictor, mediator or outcome.

In order to verify these formulations of allostatic load, we first compared associations with SES, and found that both higher sum-score and IRT score formulations of allostatic load were significantly related to lower SES. Further, in line with existing research [37,38], both formulations of allostatic load were significantly positively associated with depressive symptoms, but only the IRT score was also significantly associated with the impact of depressive symptoms on daily life. This suggests that the IRT score may be more finely tuned to tease out effects. This may also be due to the fact that by using the IRT score, we are able to

Table 3
Adjusted associations of allostatic load IRT and sum-scores with impact of depressive symptoms on daily life in the NHANES 2015–2016 study. Impact of depressive symptoms on daily life was measured by the question, "How difficult have these problems [PHQ9] made it for you to do your work, take care of things at home, or get along with people?" The response was coded as binary (not difficult at all or somewhat difficult vs. very or extremely difficult). We used logistic regression adjusted for age, sex, race/ethnicity and family income-to-poverty ratio. The models have different sample sizes, as IRT does not require complete data on all allostatic load biomarkers in order to calculate scores, unlike the sum-score approach.

Predictors	Odds Ratios	95% CI	p	Odds Ratios	95% CI	p
Intercept	0.08	0.03–0.21	**<0.001**	0.05	0.02–0.15	**<0.001**
AL IRT score	1.33	1.01–1.75	**0.045**			
AL Sum score				1.07	0.96–1.19	0.204
Sex						
Male	Reference					
Female	1.28	0.83–1.99	0.267	1.16	0.72–1.86	0.544
Age (years)	1.02	1.00–1.04	**0.041**	1.02	1.00–1.04	**0.047**
Race/Ethnicity						
Non-Hispanic White	Reference					
Mexican American	0.16	0.06–0.37	**<0.001**	0.18	0.06–0.43	**<0.001**
Other Hispanic	0.68	0.35–1.26	0.24	0.84	0.41–1.60	0.606
Non-Hispanic Black	0.46	0.25–0.81	**0.008**	0.56	0.29–1.04	0.075
Non-Hispanic Asian	0.51	0.20–1.11	0.114	0.56	0.21–1.28	0.202
Other Race, including Multi-Racial	0.68	0.23–1.66	0.438	0.72	0.21–1.90	0.546
Family income-to-poverty ratio	0.58	0.48–0.68	**<0.001**	0.6	0.50–0.72	**<0.001**
Observations	2066			1849		

include more participants in the analysis. Unlike the sum-score, the IRT score calculation does not require that every participant have every biomarker measured. Thus, we are better able to make use of missing data common to the field of psychoneuroendocrinology.

In this analysis, we build upon previous work using factor analytic approaches to explore the factor structure of allostatic load [17]. Previous factor analytic approaches have largely focused on determining the number of factors that comprise allostatic load, confirming uni-dimensionality or multi-dimensionality, and to a lesser extent focus on scoring allostatic load. In this paper, we use an IRT model to score allostatic load and determine which biomarkers provide more information. We treat each biomarker as dichotomous (high vs. low risk) in order to make comparisons with the traditional count-based calculation of allostatic load which involves a sum-score of high-risk biomarkers. Confirmatory factor analysis using dichotomous variables is asymptotically equivalent to the two-parameter logistic model, and both models can be considered as item factor analysis (Wirth and Edwards, 2007, Psychological Methods). However, the focus of the two methods are different – IRT focuses on scoring, which is of importance here and the broader discussion regarding allostatic load measurement.

It should be recognized that other investigations of allostatic load calculations have been conducted and not all allostatic load studies treat each biomarker independently. For example, across 23 biomarkers from the Midlife in the United States (MIDUS), a sum can be calculated such that each physiological system is represented with equal weight, though the number of biomarkers per system may not be equal [1]. Further, recent work on factor analysis has demonstrated that allostatic load biomarkers load onto a general allostatic load component, and within their respective physiological systems, indicating that there is common and unique variance among systems [17]. It is important to note, however, that the current analysis is limited to the number of biomarkers provided by NHANES sampling. As such, we encourage others to replicate our approach using other databases (e.g., MIDUS) with additional biomarkers.

Due to mathematical properties of the sum-score, the greater the number of biomarkers measured, the more stable the measure. However, measuring a large set of biomarkers is often infeasible, due to increased costs and participant burden. A key advantage of IRT methods is that we only need to calibrate the allostatic load scale once, and these "items" (biomarkers) can be used in future analyses to calculate an underlying allostatic load measure for a participant using the item parameter estimates, even if only a subset of biomarkers are measured, which can be used to advance reproducible research on allostatic load. However, the accuracy of a score calculated with just a small number of biomarkers may be poor. Further, this does not account for lab or instrument differences for biomarker measures. In addition, important sex, age, and race/ethnic variations in biomarkers may need to be considered as we move towards applying population norms when calculating allostatic load.

In future work, IRT may also provide a way to harmonize allostatic load scores across cohorts. Cohorts may measure slightly different sets of allostatic load biomarkers, with a common set of overlapping ones. Using IRT, we can standardize allostatic load scales across studies, using the overlapping biomarkers as anchors, so that the allostatic load score can be compared, even if the studies did not measure exactly the same set of biomarkers. IRT has been used in data harmonization, such as harmonizing measures of cognitive aging across international surveys [39], and harmonizing measures of general health functioning [40].

In this paper, we follow theoretical and methodological work that represents allostatic load as an uni-dimensional construct. Commonly used methods of calculating allostatic load scores, such the sum-score, or average biomarker z-scores, implicitly assumes that allostatic load is uni-dimensional [6,20,41]. In the interests of refining alternative measurement, our goal here was to evaluate this common sum-score approach. However, further work is needed to assess whether a multi-dimensional item response theory model better fits the data. There is potential to fit a three-dimensional IRT model (for immune, metabolic and cardiovascular physiological systems). Recent work [42] suggests that for dichotomous items, the estimated theta (allostatic load burden) scores are unbiased by violations of uni-dimensionality, that the item parameters are robust against uni-dimensionality, suggesting that practically our approach is valid even when uni-dimensionality assumptions are violated.

As stated above, we were limited by the NHANES biomarkers as we did not have the recommended three indicators for each factor (immune only has two biomarkers). Also, we did not represent neuroendocrine parameters like the HPA-axis that are central to allostatic load theory [43]. Interpretation of a multi-dimensional IRT (e.g., neuroendocrine, immune, metabolic, cardiovascular) is more complex analytically. However, multi-dimensional IRT may be helpful when we have a larger set of allostatic load biomarkers, with imbalance in the number of biomarkers per physiological system. Using multi-dimensional IRT would allow us to calculate a subscore for each physiological system, and an overall allostatic load score.

In future work, we aim to expand our IRT approach to categorical (ordinal) data, using models such as graded response models. This allows for quantiles of each allostatic load biomarker rather than a binary

measure, which will address the fact that there may not be a single threshold, but instead a gradation of risk, and this could be more reflective of subclinical risk.

4.1. Limitations

Our study had limitations. NHANES is a cross-sectional study, meaning we were unable to assess temporality of the allostatic load time course. We did not use more than one cycle of NHANES because one allostatic load biomarker, high sensitivity CRP, was only measured in this cycle, and we felt it was important to include it because we only had two immune biomarkers. NHANES is also limited in the number of allostatic load biomarkers collected. While we included the ones that are most commonly used in NHANES analyses of allostatic load [3,44], it is possible that other cohort studies may have additional measures not studied here, and the inclusion of those biomarkers in an IRT model may cause the discrimination power of the biomarkers to change. However, the use of NHANES data is a strength because we were able to calculate survey-weighted cutoffs for high-risk for each biomarker. This allows us to use cutoffs that are generalizable to the United States population, unlike other cohort studies of allostatic load in which sample-based cutoffs are used and thus findings are dependent on the sample characteristics. In future work, using clinical cutoffs as previously proposed [4, 45] can also be explored; however, this can be limited by differences in assay, machines and laboratory standards.

In future work, we will also evaluate whether it is conceptually valid to use the same scoring for all participants, or if different biomarker cutoffs should be used for different groups, to reflect physiological differences due to age, sex, race/ethnicity and comorbidity status. This may also be addressed by evaluating differential item functioning (DIF) of each biomarker [46,47]. If a biomarker exhibits DIF, this implies that given the same level of allostatic load burden, one group is more likely to score high-risk for that biomarker than another group, which suggests that different cutoffs may need to be used for different groups. This will help us define how to make an allostatic index that incorporates a balance of informative biomarkers across all race/ethnicity groups. More work is needed to study alternative methods of calculating allostatic load scores which can account for discriminating abilities of individual biomarkers, explore if different cutoffs are needed for different demographic groups, as well as assessing how these indicators function for different race/ethnicity groups.

We did not address medications in the calculation of allostatic load. While prescription medication information is available in NHANES via self-report [48] (2020), we do not know whether all participants elected to disclose their medications list or whether there is nonresponse bias. Further, many medications, including anti-hypertensives and statins, can affect biomarker levels; thus, it is difficult to ascertain which medications we should adjust for. As the focus of this paper is to demonstrate feasibility of the IRT method, we did not adjust for medications here or use it as exclusion criteria. Future work is needed to assess this IRT method in cohorts with a broader range of biomarkers and physician verified medications list.

4.2. Conclusion

In this methodological paper, we have demonstrated that IRT is able to provide additional insight into the allostatic load construct beyond that provided by the standard sum-score metric. An IRT-based approach is able to capture more variability in the allostatic load construct, as the IRT score can account for the patterns of item responses and the discriminating power of each biomarker. Because an IRT approach to calculating allostatic load appears to provide more variability in the allostatic load index than the traditional sum-score approach, this approach may be especially helpful in study designs that seek to delineate additional effects not seen using sum-scores if allostatic load is used as a predictor, mediator or outcome. Lastly, we have included a tutorial and R code in the Supplementary Materials so that researchers can calculate IRT scores in their own datasets.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Drs. Liu and Spicer are supported by the National Institute of Environmental Health Sciences (P30ES023515). Dr. Spicer is also supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R00 HD07966802). Dr. Juster is supported by Fonds de recherche Québec – Santé and holds a Canadian Institutes of Health Research Sex and Gender Science Chair.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cpnec.2020.100025.

References

[1] T.L. Gruenewald, A.S. Karlamangla, P. Hu, S. Stein-Merkin, C. Crandall, B. Koretz, T.E. Seeman, History of socioeconomic disadvantage and allostatic load in later life, Soc. Sci. Med. 74 (2012) 75–83.
[2] R.P. Juster, N.G. Smith, E. Ouellet, S. Sindi, S.J. Lupien, Sexual orientation and disclosure in relation to psychiatric symptoms, diurnal cortisol, and allostatic load, Psychosom. Med. 75 (2013) 103–116.
[3] V.M. Mays, R.P. Juster, T.J. Williamson, T.E. Seeman, S.D. Cochran, Chronic physiologic effects of stress among lesbian, gay, and bisexual adults: results from the National Health and Nutrition Examination Survey, Psychosom. Med. 80 (2018) 551–563.
[4] R.P. Juster, S. Sindi, M.F. Marin, A. Perna, A. Hashemi, J.C. Pruessner, S.J. Lupien, A clinical allostatic load index is associated with burnout symptoms and hypocortisolemic profiles in healthy workers, Psychoneuroendocrino 36 (2011) 797–805.
[5] P. Schnorpfeil, A. Noll, R. Schulze, U. Ehlert, K. Frey, J.E. Fischer, Allostatic load and work conditions, Soc. Sci. Med. 57 (2003) 647–656.
[6] T.E. Seeman, B.S. McEwen, J.W. Rowe, B.H. Singer, Allostatic load as a marker of cumulative biological risk: MacArthur studies of successful aging, P Natl Acad Sci USA 98 (2001) 4770–4775.
[7] M.E. Wallace, E.W. Harville, Allostatic load and birth outcomes among white and Black women in new orleans, Matern. Child Health J. 17 (2013) 1025–1029.
[8] R.P. Juster, M. Sasseville, C.E. Giguere, S.J. Lupien, S. Consortium, Elevated allostatic load in individuals presenting at psychiatric emergency services, J. Psychosom. Res. 115 (2018) 101–109.
[9] E. Seeman, B.H. Singer, J. Rowe, R.I. Horwitz, B. McEwen, Price of adaptation - allostatic load and its health consequences, Arch. Intern. Med. 157 (1997) 2259–2268.
[10] P. Lindfors, O. Lundberg, U. Lundberg, Allostatic load and clinical risk as related to sense of coherence in middle-aged women, Psychosom. Med. 68 (2006) 801–807.
[11] E.M. Maloney, R. Boneva, U.M. Nater, W.C. Reeves, Chronic fatigue syndrome and high allostatic load: results from a population-based case-control study in Georgia, Psychosom. Med. 71 (2009) 549–556.
[12] J. Maselko, L. Kubzansky, I. Kawachi, T. Seeman, L. Berkman, Religious service attendance and allostatic load among high-functioning elderly, Psychosom. Med. 69 (2007) 464–472.
[13] T.E. Seeman, B.H. Singer, C.D. Ryff, G. Dienberg Love, L. Levy-Storms, Social relationships, gender, and allostatic load across two age cohorts, Psychosom. Med. 64 (2002) 395–406.
[14] D.A. Glei, N. Goldman, Y.L. Chuang, M. Weinstein, Do chronic stressors lead to physiological dysregulation? Testing the theory of allostatic load, Psychosom. Med. 69 (2007) 769–776.
[15] E. Goodman, B.S. McEwen, B. Huang, L.M. Dolan, N.E. Adler, Social inequalities in biomarkers of cardiovascular risk in adolescence, Psychosom. Med. 67 (2005) 9–15.
[16] A.S. Karlamangla, B.H. Singer, T.E. Seeman, Reduction in allostatic load in older adults is associated with lower all-cause mortality risk: MacArthur studies of successful aging, Psychosom. Med. 68 (2006) 500–507.
[17] J.F. Wiley, T.L. Gruenewald, A.S. Karlamangla, T.E. Seeman, Modeling multisystem physiological dysregulation, Psychosom. Med. 78 (3) (2016 Apr) 290–301.
[18] E. Chen, G.E. Miller, M.E. Lachman, T.L. Gruenewald, T.E. Seeman, Protective factors for adults from low-childhood socioeconomic circumstances: the benefits of shift-and-persist for allostatic load, Psychosom. Med. 74 (2012) 178–186.

[19] C.L. Seplaki, N. Goldman, D. Glei, M. Weinstein, A comparative analysis of measurement approaches for physiological dysregulation in an older population, Exp. Gerontol. 40 (2005a) 438–449.
[20] M.T. Duong, B.A. Bingham, P.C. Aldana, S.T. Chung, A.E. Sumner, Variation in the calculation of allostatic load score: 21 examples from NHANES, J Racial Ethn Health Disparities 4 (2017) 455–461.
[21] X. Chen, S. Redline, A.E. Shields, D.R. Williams, M.A. Williams, Associations of allostatic load with sleep apnea, insomnia, short sleep duration, and other sleep disturbances: findings from the National Health and Nutrition Examination Survey 2005 to 2008, Ann. Epidemiol. 24 (2014) 612–619.
[22] L.N. Borrell, F.J. Dallo, N. Nguyen, Racial/ethnic disparities in all-cause mortality in U.S. adults: the effect of allostatic load, Publ. Health Rep. 125 (2010) 810–816.
[23] R.W. Kobrosly, C.L. Seplaki, C.M. Jones, E. van Wijngaarden, Physiologic dysfunction scores and cognitive function test performance in U.S. adults, Psychosom. Med. 74 (2012) 81–88.
[24] S.L. Szanton, J.K. Allen, C.L. Seplaki, K. Bandeen-Roche, L.P. Fried, Allostatic load and frailty in the women's health and aging studies, Biol. Res. Nurs. 10 (2009) 248–256.
[25] J.F. Fries, B. Bruce, D. Cella, The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes, Clin. Exp. Rheumatol. 23 (2005) S53–S57.
[26] T.H. Nguyen, H.R. Han, M.T. Kim, K.S. Chan, An introduction to item response theory for patient-reported outcome measurement, Patient 7 (2014) 23–35.
[27] G. Zipf, M. Chiappa, K.S. Porter, Y. Ostchega, B.G. Lewis, J. Dostal, National health and nutrition examination survey: plan and operations, 1999-2010, Vital Health Stat 56 (2013) 1–37.
[28] R.P. Juster, B. McEwen, S.J. Lupien, Allostatic load biomarkers of chronic stress and impact on health and cognition, Neurosci. Biobehav. Rev. 35 (2010) 2–16.
[29] C.J. Endres, nhanesA: NHANES Data Retrieval, 0.6.5, CRAN R-project, 2018.
[30] T. Lumley, Survey: Analysis of Complex Survey Samples, 3.36, CRAN R-project, 2019.
[31] D. Rizopoulos, Ltm: Latent Trait Models under IRT, CRAN R-Project, 2018.
[32] K. Kroenke, R.L. Spitzer, The PHQ-9: a new depression and diagnostic severity measure, Psychiatr. Ann. 32 (2002) 509–521.
[33] K. Kroenke, R.L. Spitzer, J.B. William, The PHQ-9: validity of a brief depression severity measure, J. Gen. Intern. Med. 16 (2001) 1606–1613.
[34] L. Khaodhiar, P.R. Ling, G.L. Blackburn, B.R. Bistrian, Serum levels of interleukin-6 and C-reactive protein correlate with body mass index across the broad range of obesity, Jpen-Parenter Enter 28 (2004) 410–415.
[35] N. Pannacciulli, F.P. Cantatore, A. Minenna, M. Bellacicco, R. Giorgino, G. De Pergola, C-reactive protein is independently associated with total body fat, central fat, and insulin resistance in adult women, Int. J. Obes. 25 (2001) 1416–1420.
[36] E.S. Rawson, P.S. Freedson, S.K. Osganian, C.E. Matthews, G. Reed, I.S. Ockene, Body mass index, but not physical activity, is associated with C-reactive protein, Med. Sci. Sports Exerc. 35 (2003) 1160–1166.
[37] R.W. Kobrosly, E. van Wijngaarden, C.L. Seplaki, D.A. Cory-Slechta, J. Moynihan, Depressive symptoms are associated with allostatic load among community-dwelling older adults, Physiol. Behav. 123 (2014) 223–230.
[38] B. McEwen, Mood disorders and allostatic load, Biol. Psychiatr. 54 (2003) 200–207.
[39] K.S. Chan, A.L. Gross, L.E. Pezzin, J. Brandt, J.D. Kasper, Harmonizing measures of cognitive performance across international surveys of aging using item response theory, J. Aging Health 27 (2015) 1392–1414.
[40] R.D. Gibbons, M.C. Perraillon, J.B. Kim, Item response theory approaches to harmonization and research synthesis, Health Serv. Outcome Res. Methodol. 14 (2014) 213–231.
[41] W.J. van der Linden, R.K. Hambleton, Handbook of Modern Item Response Theory, Springer, 1996.
[42] D.R. Crisan, J.N. Tendeiro, R.R. Meijer, Investigating the practical consequences of model misfit in unidimensional IRT models, Appl. Psychol. Meas. 41 (2017) 439–455.
[43] B. McEwen, Sex, stress and the hippocampus: allostasis, allostatic load and the aging process, Neurobiol. Aging 23 (2002) 921–939.
[44] E.M. Crimmins, M. Johnston, M. Hayward, T. Seeman, Age differences in allostatic load: an index of physiological dysregulation, Exp. Gerontol. 38 (2003) 731–734.
[45] C.L. Seplaki, N. Goldman, D. Glei, M. Weinstein, A comparative analysis of measurement approaches for physiological dysregulation in an older population, Exp. Gerontol. 40 (2005b) 438–449.
[46] M.O. Edelen, D. Thissen, J.A. Teresi, M. Kleinman, K. Ocepek-Welikson, Identification of differential item functioning using item reponses theory and the likelihood-based model comparison approach: applicaton to the Mini-Mental State Examination, Med. Care 44 (2006). S134-142.
[47] M. Zieky, History and development of DIF, in: P.W. Holland, H. Wainer (Eds.), Differential Item Functioning, Elibaum, Hillsdale, NJ, 1993.
[48] C.f.D.C.a. Prevention, National Health and Nutrition Examination Survey, Dietary Supplement and Prescription Medication Section, 2020.