# Some thoughts on composite variables

Organizing Committee of the September 20 meeting on composite variables (Sigurd Mikkelsen, Lau C Thygesen, Katrine Strandberg-Larsen, Jens Peter Bonde, Johan Hviid Andersen)

A composite variable is a variable constructed from two or more variables, usually as an unweighted sum or by multiplication or division. Composite variables are common in medical research and clinical practice.

## *Examples of composite variables*

1. Physiological indices

The body mass index (weight (kg)/(height (m))$^2$) and the waist-hip ratio (body circumference at the waist divided by the circumference at the hip) are used as measures of overweight. The Tiffeneau index (forced expiratory volume in 1 second (liter) / forced vital capacity (liter)) is used as a measure of lung function. The "face validity" rationale for creating these indices is that the nominator and the denominator are correlated and the researcher or clinician wants a measure of the nominator which is independent of the denominator (adjusted for the effects of the denominator on the nominator). Any effects of these indices could be considered a "joint effect" of the nominator and the denominator, but may be due to only one of the two exposures. Furthermore, the independent effects of the two exposures and their possible interaction are not revealed by using only a composite variable. The standard statistical way of analyzing a "joint effect" of two exposures on an outcome is a regression analysis that includes both exposures and their interaction term as covariates.

2. Indices of stress

Two dominant theories on job stress claim that stress is caused by the joint effect of two variables: In Karasek's job strain theory stress is caused by the joint effect of high demands and low control; in Siegrist's effort-reward-imbalance (ERI) theory stress is caused by the joint effect of high efforts and low rewards. In the job strain model the "joint effect" may be due to additive or interaction effects of demands and control. In the ERI model stress is caused by high efforts and by low rewards, separately, but stress caused by the joint effect of high efforts and low rewards is assumed to be higher than that of their separate effects.

Job strain is usually measured by the product of the demand scale and the control scale after a median split into binary variables. ERI is usually measured by the ratio between the effort scale and reward scale, and the effects of this ratio may be examined as a continuous variable and by categories, including dichotomization, often at unity. The rationale for creating the job strain variable and the ERI-index is the hypothesis that stress is caused by the joint action of the two component variables, but these composite variables do not reveal if an effect is due to only one of the two constituent variables. The effect of a product term of two variables or a ratio between two variables implies that the effect of either of the two variables is modified by the effect of the other variable, because the unit of the composite variable varies with both variables. Thus, whether it is stated explicitly or rejected by proponents of the theories, the use of their composite stress variables without controlling for the main effects, implies an interaction between the two component variables.

Another major stress theory is the allostatic load theory. The theory claims that chronic stress causes physiological changes (biomarkers of stress), leading to disease, e.g. heart disease. Such biomarkers could be cortisol, dehydroepiandrosterone-sulphate, C-reactive protein, fibrinogen, insulin, glycosylated hemoglobin, albumin, creatinine, pancreatic amylase, total cholesterol, high density lipoprotein cholesterol, triglycerides, systolic and diastolic blood pressure and obesity. The theory argues that, depending on the individual, the causal path from stress to a specific disease may differ between persons. One person may become obese, one develops high blood pressure, one high density lipoprotein cholesterol, one a high level of C-reactive protein, etc. Thus, stress influences stress-related biomarkers differently, depending on individual characteristics, but changes in any of these biomarkers are risk factors for stress-related disease. One person may react to stress by changes in one or more stress-related biomarkers.

In allostatic load research a set of stress-related biomarkers may be combined to a single variable (allostatic load index) indicating stress-related risk of disease, for example by dichotomizing each of the component variables at their 75th percentile in a normal population and summing these binary variables to an index ranging from 0 to the number of variables included in the index (e.g. Juster et al. Psychoneuroendocrinology 2011; 36:797—805). Such a combined set of 0/1 variables will explain less of the variation of the outcome variable than a standard regression analysis which includes component variables as separate covariates, and gives no information on the independent effect of component variables. Furthermore, to serve as a stress-index, it should be demonstrated that each component is independently associated with a "gold standard" for measuring chronic stress.

## 3. Socioeconomic status

Socioeconomic status (SES) is usually a composite variable and is often constructed from measures of occupation, income, and education. A SES index, however, will explain less of the variation of the outcome variable than a standard regression analysis that includes each component variable as a covariate and gives no information on the independent effect of component variables.

## 4. Diagnoses

Diagnoses are often composite variables constructed from a combination of dichotomous or dichotomized continuous variables (e.g. if a=1 and b=1 and (c=1 or d=1 or e=1) then the diagnosis is verified), for example classical thyreotoxicosis (Graves' disease) may be diagnosed by the presence of struma and exophtalmus and one or more other typical symptoms or findings, and verified by increased levels of thyroid hormones.

## 5. Construct scales

Scales that measure specific constructs, e.g. specific types of intelligence or specific aspects of the work environment, for example demands, control, effort and reward, are composite variables, usually constructed as the sum or mean of a number of items that are supposed to measure the construct. A number of different criteria are used to evaluate if this condition can be accepted. One condition is that the items are highly correlated, ideally with all correlations equal to 1.00, if scale items are precisely measured and properly adjusted for effects of other factors.

6. Composite outcomes

Different diagnoses, assumed to reflect the same underlying disease, may be combined, for example if there are few cases in each group of specific diagnoses (e.g. death from ischemic heart disease and a diagnosis of ischemic heart disease). See for example Freemantle et al. JAMA. 2003;289:2554-2559.

7. Physical and chemical characteristics

Measures of some physical and chemical characteristics such as speed (meter/second), acceleration (meter/second$^2$), concentration (mg/liter) and density (mass/ volume) are composite variables. In this case main effects of the components may seem theoretically meaningless.

## *Rationale for composite variables in epidemiological research?*

Are there any good reasons, a rationale, for analyzing effects of a composite variable in stead of the independent main and interaction effects of its components? And what could these reasons be?