USGS
science for a changing world

Composites:
Modeling with Formative Measures

Jim Grace

1

In this module I consider an important "third" variable type, the composite. Composites are similar to latent variables, but with some fundamentally important differences. This is a very brief introduction to the topic, so going to the primary reference below is recommended.

An appropriate citation for this material is

Grace, J.B. and Bollen, KA. 2008. Representing general theoretical concepts in structural equation models: the role of composite variables. Environmental and Ecological Statistics 15:191-213.
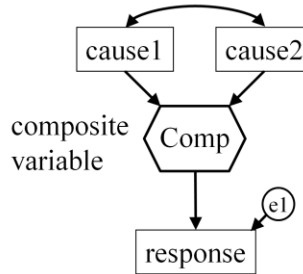
(http://www.odum.unc.edu/content/pdf/Bollen%20Grace%20Bollen%20 0(preprint%202008)%20Environ%20and%20Ecol%20Stats**.pdf**)

1. One use for composites is to represent the collective effects of a set of variables.
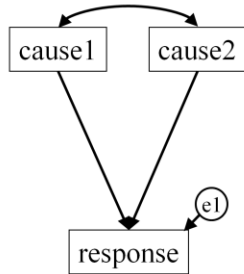
First step in compositing process.

Second step in compositing process.

Composites are a type of latent variable, but they differ from conventional LVs in some important ways. Certainly they are abstractions, but in our models their values are computed from other variables. In this example we are essentially using a composite variable "Comp" to capture the collective effects of a set of causes on some response. Thus we are translating the model on the left into the model on the right so we can represent the joint effects of cause1 and cause2 with a single path (the path from Comp to response).

In this situation we refer to the observed indicators cause1 and cause2 as "cause indicators" or as "formative indicators".

I like to use the hexagon shape to represent the composite. However, note that often it is represented in presentations by ovals, since it is a form of latent variable.

2. Composites are abstractions, and we must declare them.

Composites have an implied set of scores, similar to regular latent variables.

To specify composites in lavaan, we need to conceptualize a composite as a latent variable with no error (thus, error variance is set = 0).
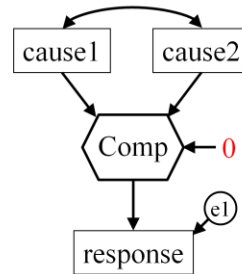
Like with regular latent variables, introducing a composite variable into a model potentially adds two parameters to our model, one for the variance of the composite and one for its scale (intercept in this case). So, we need to fix some parameters to specific values to identify our model.

≋USGS

3

Composites of the sort shown here are technically latent variables, but without variance. The absence of variance is modeled by setting the error variance to zero.

Another way of thinking about this is that the composite variable is one with a predicted set of values, one for each case in the dataset. For this model, the Comp scores are equivalent to

yhat = b0 + b1*cause1 + b2*cause2

where the bs come from the model

responses ~ lm(cause1 + cause2).

We will demonstrate this as a method of computing composites by hand later in the presentation.

Special Note: There can also exist "latent composites" with non-zero errors. Refer to "Composites with Multiple Effects" for thi case.

3. We can create composites with lavaan syntax.
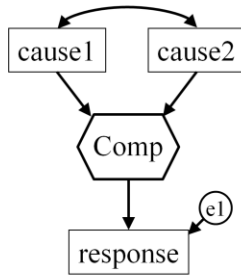

```
# we declare composite as effects of two variables

Cmod.1 <- 'Comp <~ 1*cause1 + cause2
           response ~ Comp'
```

We use "1*cause1" to set scale of the composite.

4

Lavaan has a special operator for composites "<~". Just like with latent variables, we have to give the program some information. Here we explicitly indicate that the composite is on the same scale as the first indicator by pre-multiplying cause1 by 1.

Note, this is only one of several ways composites can be specified. Refer to the illustration later in this tutorial.

3 (cont.). Creating composites with lavaan syntax - explained.

When we say "Comp <~ 1*cause1 + cause2", we are specifying one of the paths linking the composite with an indicator (in this case, "cause1").
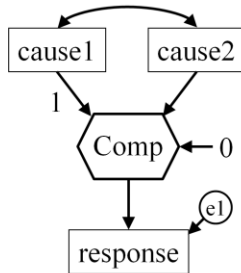
Lavaan automatically creates the variable "Comp" in this case and sets its error variance to zero.

5

Here is a little more behind-the-scenes information. This slide shows more explicitly what the syntax in the previous slide does.

Note that we could have set the scale for cause2 instead of cause1. Also, we could use a different value from 1. For example, we could use the value that came from running the model without the composite.

Further, we could set both values from causes to COMP, but to do that we would need to use the two exact values obtained from running the model without the composite (slide 2 left figure).

4. We can also compute composite scores by hand.
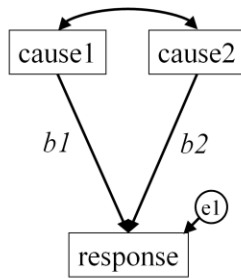
First, run model without composite and obtain coefficients.

Use coefficients to calculate composite scores.
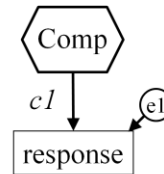
Use composite scores as a variable.

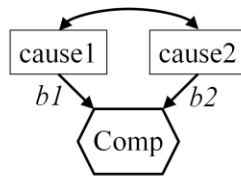The coefficient c_1 represents collective effects and is only meaningful in standardized units (normally, its raw value is 1.0).

≋USGS

It is useful to understand how to compute composite scores by hand.

One reason to do this is because lavaan can have problems solving models containing composites sometimes. Various tricks for helping lavaan include (a) premultiplying cause1 by the exact coefficient found from the model that omitted composites. In rare cases, using the composite scores compute by hand may be the only way to successfully model.

This figure indicates how we can have the values of the composite variable and model with it directly. An example of this follows.

5. An illustration of developing a composite.

≋USGS

```
## lavaan model
library(lavaan)

# model without composite (typically as a first step)
mod.1 <- 'response ~ cause1 + cause2'

mod.1.fit <- sem(mod.1, data=sim.dat)

summary(mod.1.fit, rsq= T, standardized=T)
```

	Estimate	Std.err	Z-value	P(>|z|)	Std.all
Regressions:					
response ~					
cause1	0.838	0.117	7.163	0.000	0.684
cause2	0.590	0.249	2.368	0.018	0.226
Variances:					
response	3.097	0.619			0.301
R-Square:					
response	0.699				

Note that we are going to use this value as a fixed element of our composite

7

Note that the data being analyzed are simulated from code contained in the companions file,
"SEM.10.1-Composites and Formative Indicators_code"

There are multiple ways to specify composites. Here I illustrate a slightly more advanced approach that often helps with model convergence. We note/capture the raw estimate for one of the variables to be included in the composite in the next slide.

Note also that the full output from lavaan includes two columns of standardized parameters. I have omitted the column "Std.lv", which is not relevant here as there are no "lvs" or Latent Variables in this model. The "Std.all" is the classic standardized coefficient.

If you have any questions about how to interpret coefficients, refer to the tutorial "Interpreting Coefficients" at www.nwrc.usgs.gov/SEM.

5. An illustration (cont.).

```
# lavaan model with composite
Cmod.1 <- 'Comp <~ 0.838*cause1 + cause2
          response ~ Comp'

Cmod.1.fit <- sem(Cmod.1, data=sim.dat)

summary(Cmod.1.fit, rsq=T, standardized=T)
```

	Estimate	Std.err	Z-value	P(>|z|)	Std.all
Composites:					
Comp <~					
cause1	0.838				0.818
cause2	0.590	0.305	1.937	0.053	0.270
Regressions:					
response ~					
Comp	1.000	0.140	7.163	0.000	0.836
R-Square:					
response	0.699				

interpretable coef for the composite

8

≋USGS

As stated on the previous slide, one approach to specifying the composite "Comp" here is to use the estimated raw coefficient from the previous slide. This gives us

Cmod.1 <- 'Comp <~ 0.838*cause1 + cause2

The "classic" approach to specification is to arbitrarily set one of the parameters to 1.0, for example,

Cmod.1 <- 'Comp <~ 1.0*cause1 + cause2

This is what we would have to do if we did not run the model first without the composite included.

I have a tutorial entitled "Composites – Comparing Specifications" at www.nwrc.usgs.gov/SEM that provides more depth on this topic.

5. An illustration (cont.).

```
# compute composite scores "by hand"
Comp.a <- 0.838*cause1 + 0.590*cause2

# create new data set
dat2 <- data.frame(cause1, cause2, response, Comp.a)

# model direct effect of composite on response
Cmod.2 <- 'response ~ Comp.a'

Cmod.2.fit <- sem(Cmod.2, data=dat2)

summary(Cmod.2.fit, rsq=T, standardized=T)
```

	Estimate	Std.err	Z-value	P(>|z|)	Std.all
Regressions:					
response ~					
Comp.a	1.000	0.093	10.780	0.000	0.836
R-Square:					
response	0.699				

≋USGS

we get same R-square and std. coef. as before [9]

Here we use some R code to create a composite ("Comp.a") by hand.

We then create a new dataset with the composite variable included = "dat2"

We can then model using the composite as observed variables. THIS IS ESPECIALLY HELPFUL WHEN WORKING WITH MODELS CONTAINING MORE THAN ONE COMPOSITE. IN THAT CASE, LAVAAN OFTEN HAS TROUBLE CONVERGING. THIS PROBLEM CAN BE PREVENTED BY USING COMPUTED COMPOSITE VARIABLES AND A TWO-STEP MODELING PROCEDURE.

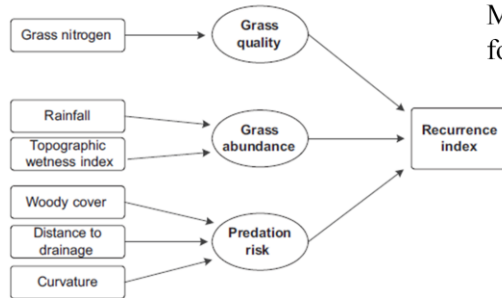6. We can use composites to represent general concepts.

Body size and the division of niche space: food and predation differentially shape the distribution of Serengeti grazers

J. Grant C. Hopcraft[1,2]*, T. Michael Anderson[3], Saleta Pérez-Vila[4], Emilian Mayemba[5] and Han Olff[1]
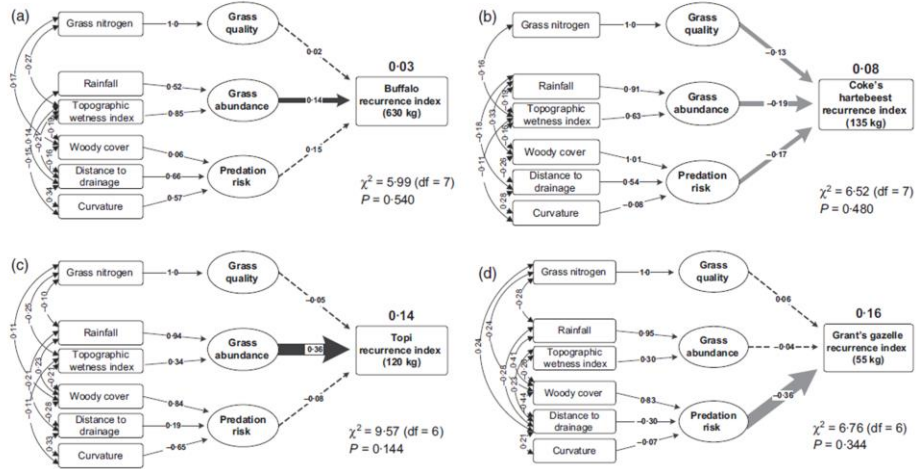
Model of predation risk versus food rewards for grazers.

Fig. 2. The *a priori* structural equation model used to assess the

USGS 10

Composites are very handy for use in addressing ecological questions.

10

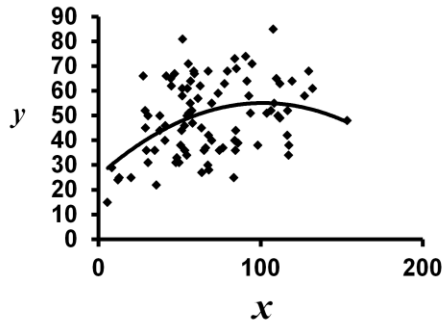7. And we can use them to compare apples and oranges (with care in interpretation).

There capacity to let us make general comparisons is one of their most appealing features.
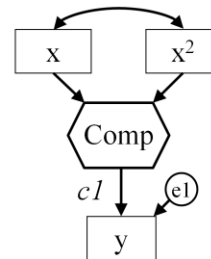
8. We can also use composites as devices.

Consider this nonlinear relationship.

We can model this using polynomial regression:

$$y = \gamma x + \gamma x^2 + \varepsilon$$

Centering a variable before squaring it reduces the autocorrelation between polynomial terms.

```
# center x before squaring
x <- x- mean(x)
x2 <- x^2
```

Note: endogenous nonlinearities covered in separate module.

12

But, composite are also handy devices, for example in polynomial modeling of nonlinear relations. What is different here is that one of the cause indicators, x-square, is really a device rather than a separate variable. Otherwise, the compositing process is similar to the case of compositing independent causes. This does not hold, however, when the nonlinearity is endogenous. In that case, the x-square term needs special consideration. This situation is covered in a separate module, "Composites for endogenous nonlinearities".