# Using R package eatDesign to compute "design" descriptives

Martin Hecht
Humboldt University Berlin, Germany

October 20, 2014

**Abstract**

In this tutorial the concepts and functionality of `eatDesign` are described and illustrated on several examples.

## 1 Terms and Concepts

A "design" is a compilation of units of several "design elements" and their relations to one another. Design elements can be thought of as relevant variables that are sources of data variation. Usually such variables are called ID variables, since they describe unique units that the data set is compiled of. The terminology and examples in this tutorial are from educational measurement research. Still, the eatDesign package can be used to calculate descriptives for all kinds of structured data. Therefore, "design" is double-quoted as it refers to just one specific term describing structured data. One package that eatDesign heavily uses and relies on is the igraph package, since for some statistics the relations between the units are first converted to a "graph". So sometimes terms from this package are used as well. For an overview of concepts in designing educational measurement studies see Frey et˜al. (2009). The first example is from this paper as well. In addition, a central concept in eatDesign is a "link" that stands for the relation between two units. So if two units are connected, they are called "linked". In igraph terminology this would mean that an edge exists that connects two vertices. One of the main purposes of eatDesign is to compute "link" descriptives for a "design".

# 2 Example 1: A Youden Square Design

In this example we use the design that is defined in Table 7 in Frey et˜al. (2009). This design is a special design called "Youden Square Design" because it possesses some special features.

```
> table7 <- data.frame ( "Booklet"  = c(1,1,2,2,3,3) ,
+                        "Position" = c(1,2,1,2,1,2) ,
+                        "Cluster"  = c(1,2,2,3,3,1) )
> table7
  Booklet Position Cluster
1       1        1       1
2       1        2       2
3       2        1       2
4       2        2       3
5       3        1       3
6       3        2       1
```

"Booklet", "Position" and "Cluster" are the design elements. The numbers in the table are the units of the design elements. Note that these units do not need to be numeric, they can be character or a factor as well. The rows of this table define the relations of the units. For instance, row 1 defines that booklet 1, position 1 and cluster 1 are connected. In terms of this concrete example, booklet 1 contains position 1, and cluster 1 is on position 1 in booklet 1.

This data frame can be used to create a "design" object using function `defineDesign`. Let's call this object `design7` in reference to `table7`.

```
> design7 <- defineDesign ( def = table7 )
```

We now have a design object that contains the design definition and descriptives that have been automatically generated. The computation of descriptives can be (very) time consuming, especially when designs get large. For some circumstances, e.g. if a lot of (partial) designs are created that are later combined (and thus, descriptives are not relevant), it might be worthwile to set `descriptives = FALSE` in `defineDesign` to save computational time. Let's have a look at the design object and its descriptives:

```
> design7
```

```
Design contains:

     3 Booklet
     2 Position
     3 Cluster

Design structure:

     Booklet and Position are completely crossed
     Booklet and Cluster are partially crossed
     Position and Cluster are completely crossed

Descriptives:

     Booklet per Position:  3
     Booklet per Cluster:   2
     Position per Cluster:  2
     Position per Booklet:  2
     Cluster per Booklet:   2
     Cluster per Position:  3

Link Descriptives:

                              linklength linkrate1 linkrate2 linkstrength linkdispersion
     Booklet linked by Position       1.00      1.00      1.00            2              0
     Booklet linked by Cluster        1.00      1.00      0.33            2              0
     Position linked by Cluster       1.00      1.00      1.00            1              0
     Position linked by Booklet       1.00      1.00      1.00            1              0
     Cluster linked by Booklet        1.00      1.00      0.33            2              0
     Cluster linked by Position       1.00      1.00      1.00            2              0

Variance-Covariance Matrix:

              Booklet Position Cluster
     Booklet     0.80     0.00    0.20
     Position    0.00     0.30    0.00
     Cluster     0.20     0.00    0.80

Design Descriptives:

     D-optimality index 1.13
     positionBalance     100
     clusterPairBalance  100
```

## 2.1 Design structure

The design structure describes the pairwise relation of design elements. Design elements can be unconnected, equivalent, nested, partially crossed or completely crossed. Unconnected elements have no "intersection" of units, that means that in the definition each unit of one element is combined with `NA` on the other element. For equivalent elements each unit of one element is uniquely combined with one unit of the other element. In this case, these

two elements are practically the same, just the labels are different. One element is nested within the other, if each unit of the element is connected to only one unit of the other element. Rather hard to label is the relationship "the other way around", so what should the relation of the element that the units of the other element are nested within be called? In eatDesign we call this "nestor", surely acknowleding that this is not the most awesome label. This relation is also not shown in the design descriptives, but if you extract the structure data frame from the object, e.g. `design7@structure`, it can eventually pop up. Completely crossed means that each unit of one element is combined with all units of the other element. If elements are partially crossed at least one unit of one element is not combined with all elements of the other element and at least one unit of one element is combined with at least two units of the other element. In short, partially crossed is "somewhere between unconnected and completely crossed". Further, being "nested" can be considered a special case of "partially crossed".

In `design7` booklets and positions are completely crossed. This means that each booklet contains all 2 positions and that each position is in every booklet. This is pretty standard and not surprising, since booklets usually should contain the same number positions and since position is defined in reference to the booklet ("Position" is actually an abbreviation for "Booklet Position") it is totally obvious that all positions are in every booklet. Booklets and clusters are partially crossed. So some – but not all – clusters are in every booklet. Positions and clusters are completely crossed, meaning that each cluster occurs at every position.

## 2.2 Descriptives

Section "Descriptives" contains basic information on the number of units of one element with reference to units of another element. If this number is not constant across units of the other element, additional statistics range, mean, median and standard deviation are displayed. In `design7` there are 3 booklets per position, 3 booklets per cluster, 2 positions per cluster, 2 positions per booklet, 2 clusters per booklet and 3 clusters per position. Note that these numbers are constant over units of the "second" element; therefore no range, mean, median and standard deviation is displayed. These descriptives are only displayed for connected elements (elements that are either completely or partially crossed, or nested). For unconnected or equivalent elements,

or if one element is nestor of the other element, these descriptives are not displayed, because they are less meaningfull and always the same (either 0 or 1, respectively). Still, they are computed and are available by extracting the descriptives slot from the object, e.g. `design7@descriptives`.

## 2.3 Link Descriptives

In section "Link Descriptives" several statistics on the magnitude of the link are displayed. Each element can be linked with respect to another element. This might or might not be relevant or interpretationally feasible in specific contexts. The link statistics have a great conceptual resemblance to the graph theory framework (which package igraph is based on). In fact, these link statistics are either just renamed, slightly adapted or aggregated statistics from igraph.

The *link length* is the *average path length*. This statistic describes how distant the units are on average. If link length is 1 (like in `design7`) each unit is directly connected to every other unit. *Link rate 1* describes the relative frequency of realized (unique) pairwise links between units in reference to all possible pairwise links of these units. In `design7` link rate 1 is 1.00 (or 100%) since all units are combined with each other. *Link rate 2* describes the relative frequency of realized pairwise links in reference to all theoretically possible pairwise links if elements were completely crossed. In `design7` link rate 2 differs over elements. For instance, link rate 2 of clusters linked by position is 1.00 because positions and clusters are completely crossed (as is shown in section design structure). Contrarily, the link rate of clusters linked by booklets are less than 100%, because booklets and clusters are only partially crossed. 0.33 (or 33%) means that there are only one third of links between units of the respective elements realized in the design compared to what would be possible if these two elements were completely crossed. The *link strength* is the average number of each unit being connected with other units. In igraph terminology, its the mean of the *degree*. So for example, if link strength is 2, on average each unit is connected to 2 other units. While link strength is the mean, *link dispersion* is the standard deviation of the degree. If link dispersion is 0 (as in `design7`), then each and every unit is connected to the number of units that is denoted by link strength.

The displayed number of digits after the decimal point is 2 by default,

except for link strength and link dispersion for which no digits are shown if they are whole-number. Full numbers (with more than 2 digits) can be accessed by extracting the link slot of the object, e.g. `design7@link`.

## 2.4 Visualization

As mentioned, the descriptives are calculated on previously defined *graph* objects. These objects are accessible in the "linkList" slot of the design object, e.g. `design7@linkList` As the name indicates linkList is a list. As the name does not indicate its a list of *graph* objects. That means we can apply all functions that are defined for graph objects, e.g. plot. Let's first have a look at the names of the linkList:

```
> names(design7@linkList)
[1] "Booklet|Position" "Booklet|Cluster"  "Position|Cluster" "Position|Booklet"
[5] "Cluster|Booklet"  "Cluster|Position"
```

The names contain two element names concatenated by "|" which stands for "linked by" as in the link descriptives output. We now can plot the units and their links for one element linked by another, e.g. clusters linked by booklets.

```
> plot(design7@linkList[["Cluster|Booklet"]])
```

In Figure 1 we see that there are three clusters (circles with numbers 1, 2 and 3). These three clusters are connected by lines that are visual displays of what we have been calling "links". As can be easily seen each cluster is connected to 2 other clusters (resulting in a link strength of 2 and link dispersion of 0, see subsection 2.3). Also, each cluster is linked to every other cluster (link rate $1 = 100\%$). Such plots will certainly not win a beauty contest, but might prove usefull for a visual inspection of the design. Besides plots a vast variety of functions that compute statistics on the graph are available in the igraph package.

## 2.5 Variance-Covariance Matrix and Design Descriptives

For interpretation of the *variance-covariance matrix* and the *D-optimality index* of the design see Frey et˜al. (2009, p. 48-49).

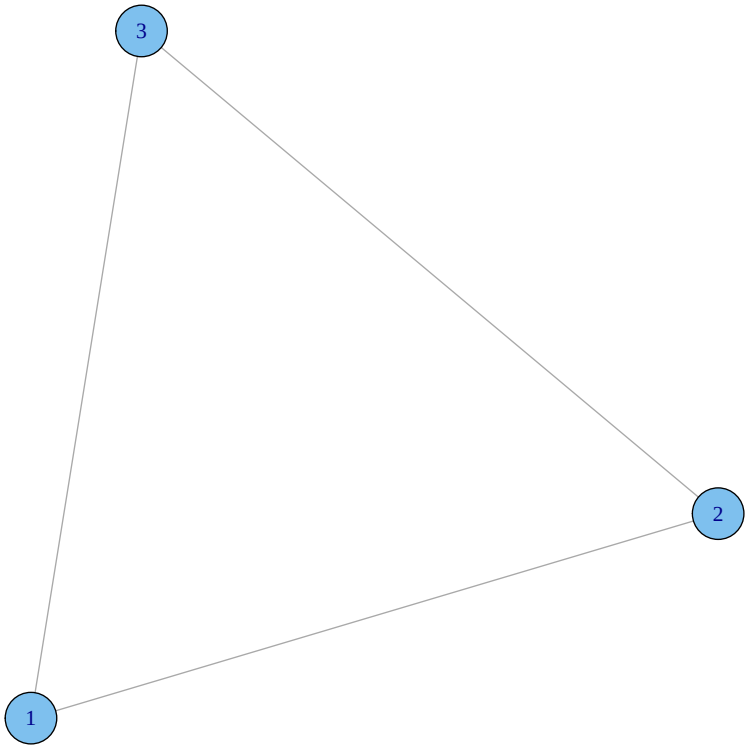


Figure 1: Links between clusters due to booklets.

# 3 Example 2: Adding one more element (items) to design7 of example 1

In example 1 we have defined 3 booklets that contain 2 positions and 3 clusters that occur on these positions in these booklets. We now want to add items to the design. Items are often combined to form clusters, so – since each item is in just one cluster – items are *nested* within clusters. Items can

be added to `design7` from example 1 in three easy steps.

First, we create a data frame that contains the items and there belonging to clusters.

```
> itemsdef <- data.frame (
+               "Item"    = paste("item",formatC(1:21,format="fg",width=2,flag="0"),sep=""),
+               "Cluster" = c ( rep(1,7), rep(2,8), rep(3,6) ) )
> itemsdef
     Item Cluster
1  item01       1
2  item02       1
3  item03       1
4  item04       1
5  item05       1
6  item06       1
7  item07       1
8  item08       2
9  item09       2
10 item10       2
11 item11       2
12 item12       2
13 item13       2
14 item14       2
15 item15       2
16 item16       3
17 item17       3
18 item18       3
19 item19       3
20 item20       3
21 item21       3
```

Note that "Cluster" must have the same unit labels as in `design7` as we want to integrate (merge) the items into `design7`. The units of "Item" can be named as favoured, however.

Second, we define a new design object containing the item-cluster relationship definition `itemsdef`.

```
> items <- defineDesign ( def = itemsdef )
> items
Design contains:

     21 Item
     3 Cluster

Design structure:

     Item are nested within Cluster
```

```
Descriptives:

     Item per Cluster:  6 - 8   M = 7.00  Mdn = 7.00  SD = 1.00

Link Descriptives:

                           linklength linkrate1 linkrate2 linkstrength linkdispersion
     Item linked by Cluster       1.00      0.30      0.10         6.10           0.83

Variance-Covariance Matrix:

              Item Cluster
     Item    38.50    4.70
     Cluster  4.70    0.65

Design Descriptives:

     D-optimality index 16.54
```

A look on the design descriptives reveals some not surprising but interesting facts. There are 21 items and 3 clusters with – as intended – items being nested within clusters. Further, each cluster contains 6 – 8, on average $M = 7.00$ items. An interesting and maybe sometimes misleading feature of the link length occurs here due to the nestedness of items in clusters. The link length is 1.00, so each unit is linked to every other unit *within* one cluster. Since items are nested within clusters, there is no *between* clusters link of items. So a link length of 1.00 does not per se mean that all units are linked to all other, as might be incorrectly assumed. Instead, this can actually be seen by link rate 1 which is 0.30 and thus below 1.00. Again, we can look at the plot for further insight.

```
> plot(items@linkList[["Item|Cluster"]])
```

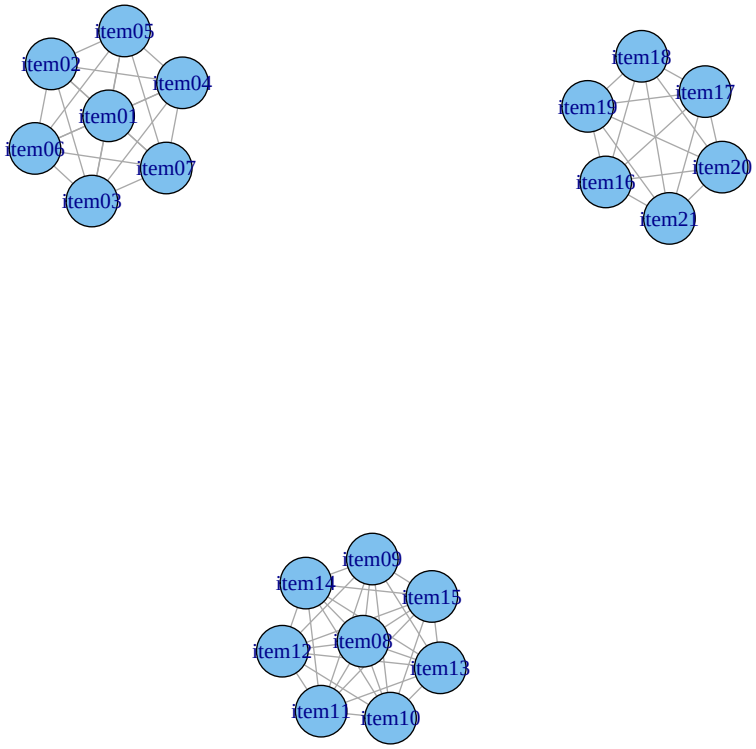


Figure 2: Links between items due to clusters.

Third, we add the new design `items` to the original design `design7` using the "+" operator and call the resulting design `design2`.

```
> design2 <- design7 + items
> design2
Design contains:

     3 Booklet
     2 Position
     3 Cluster
```

```
     21 Item

Design structure:

     Booklet and Position are completely crossed
     Booklet and Cluster are partially crossed
     Position and Cluster are completely crossed
     Booklet and Item are partially crossed
     Position and Item are completely crossed
     Item are nested within Cluster

Descriptives:

     Booklet per Position:  3
     Booklet per Cluster:   2
     Position per Cluster:  2
     Booklet per Item:      2
     Position per Item:     2
     Position per Booklet:  2
     Cluster per Booklet:   2
     Item per Booklet:      13 - 15   M = 14.00  Mdn = 14.00  SD = 1.00
     Cluster per Position:  3
     Item per Position:     21
     Item per Cluster:      6 - 8   M = 7.00  Mdn = 7.00  SD = 1.00

Link Descriptives:

                              linklength linkrate1 linkrate2 linkstrength linkdispersion
     Booklet linked by Position       1.00      1.00      1.00            2              0
     Booklet linked by Cluster        1.00      1.00      0.33            2              0
     Position linked by Cluster       1.00      1.00      1.00            1              0
     Booklet linked by Item           1.00      1.00      0.33            2              0
     Position linked by Item          1.00      1.00      1.00            1              0
     Position linked by Booklet       1.00      1.00      1.00            1              0
     Cluster linked by Booklet        1.00      1.00      0.33            2              0
     Item linked by Booklet           1.00      1.00      0.43           20              0
     Cluster linked by Position       1.00      1.00      1.00            2              0
     Item linked by Position          1.00      1.00      1.00           20              0
     Item linked by Cluster           1.00      0.30      0.10         6.10           0.83

Variance-Covariance Matrix:

              Booklet Position Cluster  Item
     Booklet     0.71    -0.04    0.32  2.29
     Position   -0.04     0.26    0.00  0.00
     Cluster     0.32     0.00    0.63  4.59
     Item        2.29     0.00    4.59 37.56

Design Descriptives:

     D-optimality index 24.31
     positionBalance      100
     clusterPairBalance   100
```

Probably the most important gain from this maneuver are the descriptives of the added element (items) in reference to the original elements (booklets

and positions). For instance there are 13 – 15 items per booklet and items have a link strength of 20 and link dispersion of 0 due to the linkage that is created by booklets. And again, a plot can be created, e.g. for the items that are linked by booklets.

```
> plot(design2@linkList[["Item|Booklet"]])
```

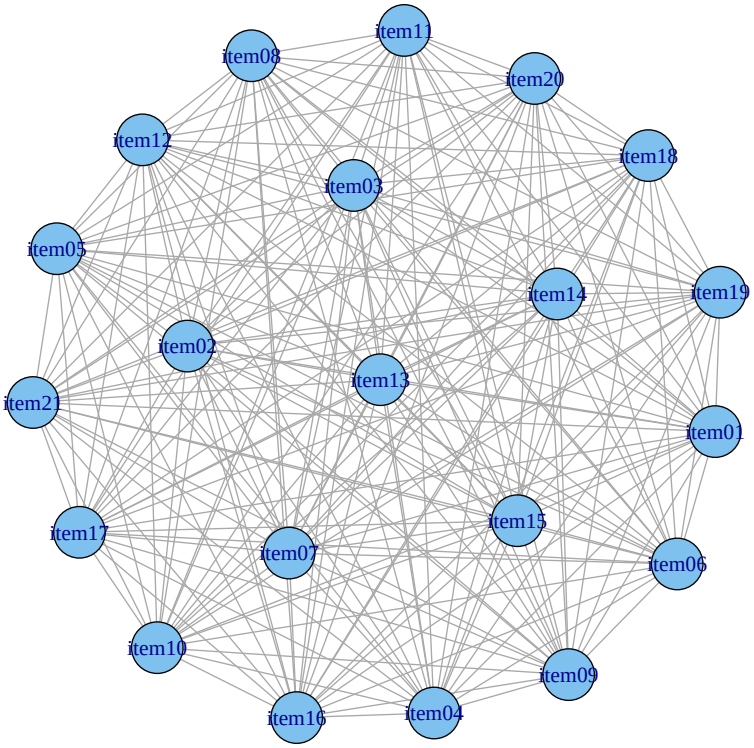


Figure 3: Links between items due to booklets.

# References

Andreas Frey, Johannes Hartig, and Andre˜A. Rupp. An NCME instructional module on booklet designs in large–scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, 28(3):39–53, 2009.