

## 2020 United States COVID-19 Vaccination Preference (CVP) Study

### Appendix 3. how to evaluate a design matrix

21 December 2020

Benjamin M. Craig, PhD, [benjamin.craig@iahpr.org](mailto:benjamin.craig@iahpr.org)

The 2020 US CVP study was originally conducted as a worked example for the forthcoming textbook, “Methods for Health Preference Research.” As requested by the reviewers of the article, this section is an excerpt from its current draft that I wrote and shows how a study team might evaluate a design matrix in terms of **coverage** after it has been assembled, generated, or optimized.

A design matrix may seem appropriate under idealized conditions; however, the study team should seek assurances that the matrix will satisfy the minimum statistical properties for estimation under the **worst-case scenarios**. Assessing a design matrix in terms of coverage may seem tedious and unnecessary, but careful evaluation prevents wasting scarce data collection resources.

The worked example has three design matrices, totaling 168 unique sets (56 x 3). The random design was assembled by randomly selecting candidate sets, the generator-developed design was generated from an orthogonal array, and the efficient design was optimized based on a conditional logit, D-error, and fixed priors  $\beta = (0, 0, 0.2, 0.2, 0.1, 0.2, 0.3)$  under the assumption of preference homogeneity. Each design was constructed to be the same size, includes overlaps on each attribute, and excludes dominated alternatives and duplicate sets. We now review how each design may be evaluated using their response matrices.

### Constructing a response matrix as an evaluative tool

The best way to evaluate a design matrix is to examine the potential responses,  $\mathbf{y}$ , that it can produce, namely the **response matrix**. To create the response matrix, a study team expands each set into its possible responses (e.g., A, B, C) and each response is exploded into its equality statements (A>B, A>C, B>A, B>C, C>A, C>B). These statements represent the potential preference evidence from a preference elicitation task with this set. A response matrix is an evaluative tool that represents all possible evidence given a design matrix, such that each row resolves an ambiguity and each column represents a possible tradeoff along a single attribute.

Mathematically, a response matrix has  $S \times J \times (J - 1)$  rows and  $K$  columns of tradeoffs. Each row represents an inequality statement corresponding to the response,  $\mathbf{y}_{sj}$ , where response,  $\mathbf{y}$ , is object  $\mathbf{j}$  of set  $\mathbf{s}$ . Each tradeoff,  $x_k$ , is a **balanced ternary variable** (+1, 0, -1), indicating whether the tradeoff favors the preferred alternative (+1), its counterpart (-1) or neither (0). Each row is like a pro-con list under Franklin’s Rule, where the choice is known, the positive tradeoffs represent the “pro,” and the negative tradeoffs represent the “con”.

These ternary variables,  $x_k$ , are **dummy-coded** to indicate the effect of the attribute levels on the choice between the two alternatives (increasing or decreasing the probability) relative to another level (typically the level below itself). For the worked example, Table 1 shows the

response matrix for a single set (11111, 21114, 22122) where the third attribute is a holdout, and the sixth tradeoff is not observed regardless of response.

Table 1. Response matrix for a single task in the worked example

$y$	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$
11111>21114	1	0	0	0	0	0	1
11111>22122	1	1	0	1	1	0	0
21114>11111	-1	0	0	0	0	0	-1
21114>22122	0	1	0	1	1	0	-1
22122>11111	-1	-1	0	-1	-1	0	0
22122>21114	0	-1	0	-1	-1	0	1

If two alternatives are identical, the response does not resolve ambiguity, so this response would be excluded from the response matrix. A common mistake is to drop duplicate rows of the response matrix prior to its evaluation. Generally, duplicates rows of  $x$  are undesirable, but tolerated. (A parallel in qualitative work is asking the same question twice with slightly different wording.) The variance of any ternary variable  $x_k$  is the frequency of the nonzero values divided by the number of rows minus one. More advanced designs may include linearity constraints or allow for order effects, replacing the main ternary variables with other balanced variables (e.g., differences in numerical cost or risk, interactions).

In the worked example, the response matrix for each design has 336 rows ( $56 \times 3 \times 2$ ) and seven columns ( $K = 7$ ), including only the seven main effects, not the opt-out. By ignoring the opt-out in the task, the three design matrices are evaluated under the assumption that opt-out is never selected. Each of the three design matrices selected its 56 sets from the 25,360 candidate sets, which are based on 64 possible objects and 1,720 candidate pairs (i.e., the 2,016 possible pairs  $[64 \times 63 / 2]$  excluding those with dominated alternatives). As shown in Table 2, the candidate pairs reduce to 499 unique combinations of tradeoffs (potential rows of the response matrix). Among these 499, 19% have no holdout—that is, no attribute shared by both alternatives in a pair—and the remainder have one or more holdouts.

Comparing the response matrices in the three designs, three differences are worth mentioning. First, the efficient design includes fewer unique objects. Second, the efficient design includes more unique *responses*. Third, the efficient design includes more responses without holdouts and fewer responses with holdouts on the fifth attribute. These differences are not necessarily good or bad, but it illustrates how different set selection approaches in experimental design create differences in the preference evidence.

Table 2 Comparison of the three design matrices in the worked example

	All	Random	Generator-developed	Efficient	Combined
Unique sets	25,360	56	56	56	168
Unique objects	64	63	62	48	64
Unique pairs	1720	150	132	150	401
Unique responses	499	59	52	113	185
Holdouts (proportion of unique combinations where an attribute level is shared)					
Attribute 1	30%	44%	35%	40%	37%
Attribute 2	30%	37%	27%	37%	35%
Attribute 3	32%	73%	58%	34%	46%
Attribute 4	32%	59%	58%	27%	41%
Attribute 5	7%	19%	15%	5%	10%
No holdouts	19%	0%	0%	12%	8%

## Frequency of tradeoffs and multicollinearity

Even if a subject is assigned all sets, the response matrix for a single respondent has a subset of the rows. For the worked example, when a person chooses A, only two rows (A>B, A>C) of the six rows (A>B, A>C, B>A, B>C, C>A, C>B) will be observed and the other four rows (B>A, B>C, C>A, C>B) will not be observed. When a person chooses A, preference between B and C is endogenously censored. This is because the comparison is made over three alternatives (four if including an opt-out), so effectively, we do not observe “second chance” preferences among the discarded alternatives.

To observe all rows, a study must include repeated tasks. If the number of responses per set is large and the design excludes dominated alternatives, the preference evidence on all tradeoffs will be observed. [1] After the initial review of unique sets and holdouts (Table 2), the unweighted response matrix is assessed based on the frequency of tradeoffs (non-zero elements) and their multicollinearity.

**Frequency of tradeoffs** is assessed by examining the proportion of nonzero elements for each tradeoff. This frequency is proportional to the variance and characterizes the amount of information on its effect. For example, if an attribute is always a hold out (i.e., frequency of zero), the design lacks the variation needed to estimate the effect of that attribute. Note that balance in frequency implies level balance (the number of times that an attribute level is shown), but level balance does not imply balance in frequency. As an extreme example, an attribute at Level 2 may occur with the same frequency as any other attribute level (level balance), but appear only as a hold out; therefore, the design does not have any preference evidence on this attribute level. When the frequency of a tradeoff is near zero (like an underpowered arm of a clinical trial), the study team may either increase the sample size or abandon the design matrix.

**Multicollinearity** is assessed by examining the pairwise relationship between ternary variables. If two variables are perfectly correlated (**perfect collinearity**), the design lacks the variation

needed to allow the estimation of their independent effects. If two variables are highly correlated (**multicollinearity**), the sample size may be too small to identify their independent effects, a problem of **micronumerosity** or insufficient power. [2] For the purposes of design evaluation, study teams may check for multicollinearity by computing the minimum and maximum correlations among the ternary variables.

By construction, tradeoffs on the same attribute will be positively correlated, and tradeoffs on different attributes have near-zero or negative correlations. As a general rule, all correlations should be between -0.7 and 0.7. Some teams may also estimate the determinants of the correlation and covariance matrices as an expedient measure of collinearity. [3] If all ternary variables were perfectly uncorrelated, the determinant of the correlation matrix equals 1. [4, 5] Maximizing the determinant of the covariance matrix is sometimes known as **D-optimality** [6] or **D-efficiency** [7], but should not be confused with the minimization of D-error, previously described.

After the assessment of the unweighted response matrix, the final component involves the assessment of the design matrix under worst-case scenarios, which entails violations of utility balance.

## Utility balance and worst-case scenarios

**Utility balance** implies observing each row of the response matrix with equal likelihood,  $P(y_{sj} = 1|X, \beta) = w_{sj} = 1/J$ . Study teams typically start by evaluating the frequencies of tradeoffs and correlations under this unrealistic scenario. After this unweighted evaluation, the study team computes the likelihood of each row  $w_{js}$  under each worst-case scenario and re-evaluates the weighted response matrices. Unlike priors, which represent the most likely parameter set, these worst-case scenarios characterize extreme parameter sets that may reasonably occur.

The weighted evaluation starts with the most likely scenario based on the priors. This evidence serves as a reference to better understand coverage under the worst-case scenarios. After the prior, we conduct **one-way analyses**, iteratively placing extreme values on each tradeoff and keeping the rest at the prior. In this analysis, any pair with the selected tradeoff is deterministic; therefore, the response matrix includes only the pairs where the attribute is a holdout. Apart from one-way analyses, the study team may assess the impact of increasing the priors by a factor of three or more, mimicking heightened preference intensity.

Regardless of scenario, the evaluative process involves computing the weights  $w_{sj}$  given a parameter set and estimating the percentage of utility balance, known as **B-error** (i.e.,  $\frac{1}{S} \sum_{s=1}^S \prod_{j=1}^J J \times w_{sj}$ ). [8] In the unweighted evaluation (prior values of zero), B-error is one by definition which implies complete utility balance. An unsatisfactory design matrix may include weights  $w_{sj}$  near zero or be highly unbalanced overall, having a B-error below 70%.

For the worked example, Table 3 shows D-, A-, and B-errors, the minimum weights, the minimum frequencies, and the correlation ranges based on five scenarios: unweighted (zero priors), priors, and two worst-case scenarios.

Table 3. Evaluation of the three design matrices in the worked example

	D-Error	A-Error	B-Error	Minimum weight	Minimum frequency	Minimum correlation	Maximum correlation
Unweighted (0,0,0,0,0,0,0)							
Random	0.173	0.212	1.000	0.333	0.262	-0.131	0.566
Generator-developed	0.137	0.176	1.000	0.333	0.286	0.000	0.577
Efficient	0.130	0.189	1.000	0.333	0.440	-0.496	0.598
Priors (0,0,0.2,0.2,0.1,0.2,0.3)							
Random	0.174	0.214	0.984	0.260	0.256	-0.131	0.572
Generator-developed	0.139	0.178	0.972	0.242	0.276	-0.015	0.577
Efficient	0.130	0.189	0.995	0.301	0.441	-0.495	0.598
One-way analysis (0,0,10,0.2,0.1,0.2,0.3)							
Random	0.224	0.290	0.986	0.260	0.316	-0.167	0.614
Generator-developed	0.241	0.299	0.980	0.250	0.323	-0.134	0.573
Efficient	0.354	0.695	0.998	0.475	0.402	-0.729	0.626
Five times the priors (0,0,1,1,0.5,1, 1.5)							
Random	0.204	0.251	0.696	0.078	0.235	-0.158	0.584
Generator-developed	0.175	0.229	0.558	0.049	0.249	-0.068	0.567
Efficient	0.139	0.202	0.877	0.186	0.444	-0.494	0.599

The one-way analysis in Table 3 mimics the scenario when the third attribute becomes deterministic, removing any mention of this tradeoff from the design. This extreme scenario has little effect on the random and generator-developed designs but causes the efficient design to appear inferior to its counterparts in terms of D- and A- error and minimum correlation. Similar results were found in the one-way analyses of the three other ordinal attributes. The last worst-case scenario demonstrates the effects of increasing the priors by a factor of five, mainly the decrease in the B-error and minimum weights. Overall, each design performs well under these assumptions. The reader is reminded that real world performance of the design depends on the true parameters, which are unknown. The best that we can do at this stage is to evaluate the design matrix under a variety of reasonable and extreme scenarios.

## Summary

The evaluation of the response matrix is useful to eliminate clearly bad design matrices, increasing the chance that the overall experimental design will perform well when data are collected. The evaluation is not meant to inform a choice between two good designs, a question of design efficiency. The benefits of switching from a satisfactory to a superb design matrix are typically unnoticeable. [9] Higher levels of refinement may matter in the cases when samples are unavoidably small (e.g., rare disease patients) or when assumptions potentially fail (e.g., McFadden’s positivity assumption, misinformed priors, unknown likelihood functions). More generally, the outcome of matrix evaluation is either pass or fail.

## References

1. McFadden D. Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka P, editor. *Frontiers in Econometrics*. Academic Press; 1974. p. 105-42.
2. Goldberger AS. *A course in econometrics*. Harvard University Press; 1991.
3. Wald A. On the efficient design of statistical investigations. *The annals of mathematical statistics*. 1943;14(2):134-40.
4. Bartlett MS. Tests of significance in factor analysis. *British journal of psychology*. 1950.
5. Jiang T. Determinant of sample correlation matrix with application. *The Annals of Applied Probability*. 2019;29(3):1356-97.
6. Kiefer J, Wolfowitz J. Optimum designs in regression problems. *The Annals of Mathematical Statistics*. 1959:271-94.
7. Kuhfeld WF, Tobias RD, Garratt M. Efficient Experimental Design with Marketing Research Applications. *Journal of Marketing Research*. 1994;31(4):545-57. doi:10.2307/3151882.
8. Scarpa R, Rose JM. Design efficiency for non-market valuation with choice modelling: how to measure it, what to report and why\*. *Australian Journal of Agricultural and Resource Economics*. 2008;52(3):253-82. doi:10.1111/j.1467-8489.2007.00436.x.
9. Norman R, Craig BM, Hansen P, Jonker MF, Rose J, Street DJ et al. Issues in the Design of Discrete Choice Experiments. *The Patient - Patient-Centered Outcomes Research*. 2019;12(3):281-5. doi:10.1007/s40271-018-0346-0.