On the Reproducibility of Latent Variables

William D Heavlin, Google, Inc, P O Box 2846, El Granada CA 94018

September 28, 2017

Abstract

Latent variables are typically constructed to explain variance and optimize validity, then later refined to enhance reproducibility. Here we promote reproducibility to a first-order property. Our example uses the ratings of NFL quarterbacks. Emphasizing reproducibility corresponds to a shift in purpose — from describing recent passer performance to predicting its future level. The NFL passer rating is a weighted sum of four ratios: the numerators consist of counts of completed passes, yardage, touchdowns, and interceptions; the denominator is the number of attempted passes. Our objective function maximizes the correlation between past and future passer ratings by the choice of weights. At our disposal are two classes of weights: (a) weights on the four ratios and (b) weights on recent and less recent games. The algorithm itself we term Spokane, for symmetric, positive canonical correlation.

key words: canonical correlation, gamuts, latent variables, Spokane, sports statistics

1 Latent Variables

We introduce latent variables (LVs) by three examples: intelligence quotients (IQ), customer engagement, and quarterback passing ratings.

Quantifying LVs has arguably advanced western civilization. For example, Spearman's (1904) original concept of intelligence, the IQ score, sought to displace pre-scientific proxies such as class accent, parentage, and dress. More recent evolutions, e.g. Gardner (1993), decompose intelligence into multiple dimensions; this refinement effort is inextricable with the LV measurement paradigm.

From this example, one sees that LVs are intellectual constructs, not directly observable. Instead, LVs are postulated to lie behind observed variables. LVs separate meaning from measurement, establishing a certain intellectual architecture. And postulating latent variables engenders problems: how to quantify LVs, and with what qualities are they to be constructed.

In the context of this paper, LVs assert properties of observational units, usually human: Each person has his own IQ, each customer relates business to a different degree, each quarter has a different skill level. In this sense, LVs differ from the hyper-parameters of Bayesian models, which represent properties of a larger data system. LVs also manifest through multiple measurements, so LVs differ from random effects, which are essentially residuals with respect to a model.

In data analysis, LVs play two roles. First, they can define *responses*. For example, one might want to measure the cognitive benefits of young-child nutritional programs; an improvement of four IQ points quantifies the result and makes the study amenable to statistical analysis. Second, LVs can define *gamuts* or stratifications. In the same experiment, one might distinguish the improvement among children with sub-100 baseline IQs from those with 100-plus baselines.

As this latter example illustrates, gamuts describe in which subsets improvements occur, a topic of persistent scientific interest. In some cases, e.g. baseline IQ, identifying the gamut is straightforward, in others less so. We refer to the problem of constructing a useful gamut the *make-a-gamut problem*.

The present research is strongly motivated by the make-a-gamut problem. Multiple Google product areas want their businesses to grow. Some growth comes from obtaining new users, the remainder from increasing the usage of existing users. In a baseline period, users are *engaged* — engagement is the key latent variable here — to different degrees. Changes to product features and interface can nudge some users toward more use, more engagement. But not all increases in engagement are equal; increases in engagement from low-engagement users are usually more valuable: Low-engagement users serve as proxies for zero-engagement users, which suggests leverage greater than the size of their own cohort. Further, nudging low-engagement users to higher levels moves them to a level from which further growth can occur. Finally, in advertising, low-engagement users are generally considered worth more — they are fresher and less saturated.

In the each of the cases of IQ, engagement, and quarterback ratings, the gamut is the result of a deliberate intellectual and statistical exercise. In practice, there are two approaches to making gamuts, supervised and unsupervised. Heavlin (2016) describes the supervised application. The present work addresses the unsupervised approach to gamut construction. Both approaches pay explicit attention to prediction error.

Historically, the empirical construction of LVs emphasizes psychometric validity — its interpretation and meaning. Secondary properties like measurement reproducibility, are treated as tasks for downstream engineering, increasing questionnaire size, administrative discipline, etc.

This work departs from the standard validity-first, reproducibility-later task order. Measuring IQ means labeling individuals with a property with of some permanence. A business that quantifies customer engagement typically formulates different strategies for low and highly engaged users; the implicit assumption is that these low and high labels remain relevant in the future. Likewise, rating quarterbacks does not make really sense unless the ratings somehow predict future performance. If such labels are intended maintain their meaning over time, then any re-assessment should be result in about the same label — in a word, the label should be reproducible. As we demonstrate below, incorporating reproducibility as a first-order property has positive benefits.

2 Quarterback Ratings

The remainder of this paper concentrates on the quarterback ratings (QBRs) of the National Football League. QBRs are arguably the LV most visible to American culture at large. This data is also less proprietary than, say, customer engagement scores; through Fantasy Football Today (2017), these scores are readily available. QBRs are also compact, aiding the exposition below.

QBRs summarize four statistics: total passes completed, total yards obtained in pass plays, touchdowns, and interceptions. After dividing by the corresponding number of attempted passes (and denominator) by ATT, denote the resulting ratios by COMP, YARD, TD, and INT, respectively.

The algorithmic calculation for QBR is as follows:

```
bound(x,theta) ≡ min( max(x,0), theta)
a <- bound(5*(COMP-.3),    theta=2.375)
b <- bound(0.25*(YARD-3), theta=2.375)
c <- bound(20*TD,          theta=2.375)
d <- bound(2.375-25*INT,   theta=2.375)
QBR <- 100*(a+b+c+d)/6
```

Note that expression `d` changes a lower-is-better `INT` ratio to a higher-is-better value. In practice, `bound` rarely imposes the 0 floor or 2.375 ceiling only rarely.

QBRs are always between 0 and 158.3. As originally proposed in 1970, average values of `a`, `b`, `c`, and `d` were 1.0, so average performance then was 66.7. Since then the NFL average has since risen, to 83 in 2012 and 2013, and 84 in the most recent season, 2016.

As Figure 1(a) illustrates, the correlation between QBR and the first principal component (PCA1, Hotelling, 1933) is quite high; the Spearman rank correlation is 0.956. This is not a coincidence. Principal components obviously guided the original 1970 construction of QBR. In this paper, we treat QBR and PCA1 as nearly synonymous, two manifestations of a single idea.

Figure 1(b) shows the point of departure for this paper. From one game to another, the QBR has an autocorrelation of 0.10 — at a basic level, QBR fails to predict. Figure 1(c) presents for autocorrelation for PCA1; its correlation is even lower, 0.088.

By its close association with PCA1, QBR maximizes the variation in performance among the quarterbacks; in a given week, QBR calls attention to the most extreme narratives of success and failure. But recent performance levels are not predictions. This is bad: it implies that a good performance promises little about next week's performance. And this is good: statistically, a quarterback with a poor performance has at least the potential to rebound the week after. Can we construct a quarterback rating that better predicts future performance?

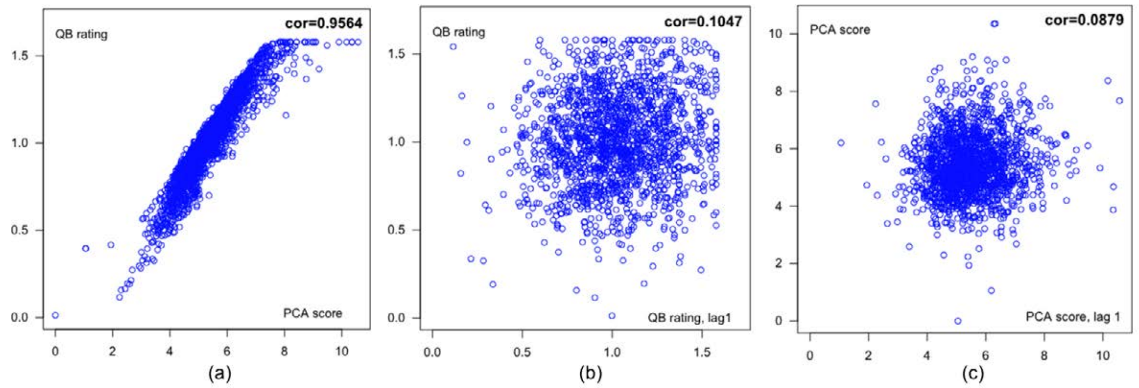

Figure 1: Scatterplots: (a) QBR and PCA1, (b) QBR and QBR lag 1, (c) PCA1 and PCA lag 1.

3 Approach

The example of quarterback ratings motivates the following project. First, we need a set of data to work with. Secondly, we need to define our criterion by which we can separate good options from bad. Thirdly, we need to define a decision space over which we optimize.

For our data, we make use of the quarterback statistics available at Fantasy Football Today (2017). In particular, we use the data from the regular season games from the five seasons 2012-2013 through 2016-2017. This choice gives relatively fresh data with suitable volume. We focus on the currently defined component ratios COMP, YARD, TD, and INT. This keeps the scope of our effort well within the range of current quarterback rating practice.

Regarding our criterion, we emphasize the ability to predict the future. Detailed below, for any candidate quarterback rating $q(t)$ at time t, we quantify how well it predicts future performance $q(t+)$, and we engineer the details in order to achieve this property.

A close reading of the previous two paragraphs reveals the two natural dimensions to our decision space: (1) The current QBR has weights (inside bound) of 5, 0.25, 20, and 25, respectively; other weights are conceivable, including weights of zero, which amounts to dropping a component. (In what follows, we do not use the function bound. This simplification restricts us to strictly linear calculations.)

(2) QBR is typically applied to a quarterback's performance on a given day. This emphasizes how the quarterback does in a particular game, with certain available receivers, in the context of a single defense. But one can imagine predicting not only the next game, but also predicting longer periods of time, the remainder of the season, say. By symmetry, one can describe a quarterback's performance not only by the most recent game, but also over the last two or 8 games. This allows quarterback ratings the statistical benefit of larger sample sizes, with the potential loss of freshness.

3.1 Notation

Fix a point in time, t_0, the *meridian*. The previous period (game) has index $t_0 - 1$ ("lag 1"), the next period has index $t_0 + 1$. The previous k periods have indices $t_0 - 1, ..., t_0 - k$.

By convention, we attach the present period t_0 to the past: *Span* $-k$ refers to periods $t_0, t_0 - 1, ..., t_0 - (k-1)$, and is denoted by $S(t_0, -k)$. Conversely, the next k time periods $t_0 + 1, t_0 + 2, ..., t_0 + k$ we denote by $S(t_0, +k)$; note that $t_0 \in S(t_0, -k)$ but $t_0 \notin S(t_0, +k)$; $S(t_0, -k)$ refers to the most recent k periods while $S(t_0, +k)$ refers to the k periods that occur after t_0.

Observational units we index by $i, i = 1, ..., n$. For a given span S, metrics are denoted as follows: $X_{i,j=1}(S)$ is COMP$[i, S]$, $X_{i,j=2}(S)$ is YARD$[i, S]$, $X_{i,j=3}(S)$ is TD$[i, S]$, and $X_{i,j=4}(S)$ is INT$[i, S]$. Metrics for unit i and span S have weights $n_i(S)$; here $n_i(S)$=ATT$[i, S]$.

By considering multiple spans, metrics readily proliferate: Note that span $S(t, +k+1) = S(t, +k) \cup (t + k + 1)$ and $S(t, -(k+1)) = S(t, -k) \cup (t - k)$. As a result, for a given meridian t and metric X, one can consider the multiple metrics $X[S(t, k)]$, $X[S(t, k+1)]$, $X[S(t, k+2)]$, ... The set of metrics $X[S(t, 1)], X[S(t, 2)], ..., X[S(t, H)]$ we call the *horizon* of X at meridian t. The horizon data structure allows us to weight together the metrics of different spans.

		current				future			
	corrs	COMP	YARD	TD	INT	COMP	YARD	TD	INT
current	COMP	1.000	0.566	0.353	-0.279	0.103	0.084	0.091	-0.061
	YARD	0.566	1.000	0.554	-0.201	0.021	0.057	0.028	-0.022
	TD	0.353	0.554	1.000	-0.216	0.032	0.069	0.054	-0.010
	INT	-0.279	-0.201	-0.216	1.000	-0.055	-0.058	-0.095	0.025
future	COMP	0.103	0.021	0.032	0.055	1.000	0.573	0.353	-0.280
	YARD	0.084	0.057	0.069	0.058	0.573	1.000	0.530	-0.213
	TD	0.091	0.028	0.054	0.095	0.353	0.530	1.000	-0.210
	INT	-0.061	-0.022	-0.010	0.025	-0.280	-0.213	-0.210	1.000

Table 1: Correlations of Span=±1 variables. Left four: variables of past, spans $S(k = -1)$. Rightmost four: variables of future, spans $S(k = +1)$. All feasible meridians combined.

3.2 Correlations

Consider Table 1. These are the correlations of the span=1 variables; the four leftmost variables denote those before the meridian, $k = -1$, the four rightmost denote those after. Note that the upper left 4×4 and lower right 4×4 sub-matrices, R_{--} and R_{++} respectively, are numerically quite similar. Matrices such as these define the principal components calculation alluded to in section 2. The first eigenvector of $R_{--} + R_{++}$ is $w_{00} \propto (1.00, 1.08, 0.95, -0.63)$.

Our particular interest is with the upper right 4×4 matrix R_{-+} of Table 1; these correlations indicate how well the variability of span=-1 predicts the variability of span=+1. Consider a weighted combination of the four metrics, $m(w) = w_1 \times \text{comp} + w_2 \times \text{yard} + w_3 \times \text{td} - w_4 \times \text{int}$ and likewise $M(w) = w_1 \times \text{COMP} + w_2 \times \text{YARD} + w_3 \times \text{TD} - w_4 \times \text{INT}$.

Table 1 gives us sufficient information for calculating the correlation between $m(w)$ and $M(w)$; indeed, the correlation of $m(w)$ and $M(w)$,

$$\mathbb{COR}(m(w), M(w)) = \frac{w^T R_{-+} w}{[w^T R_{--} w]^{1/2} [w^T R_{++} w]^{1/2}},$$

a ratio of matrix multiplications. This correlation maximized by $w_{-+} \propto (1.00, -0.19, 0.26, -0.66)$, achieving 0.115. Compare this to the 0.102 correlation achieved by w_{00}.

These two examples of $w_{00}, 0.102$ and $w_{-+}, 0.115$ encapsulate the remainder of this work: (a) We want to maximize the next-period correlation, which we call here *reproducibility.* (b) The particular solution w_{-+} is unattractive: the sign of the second component, for yardage, is negative. (c) Both achieved correlations, 0.102 and 0.115, are uncomfortably small, and we need something qualitatively beyond re-weighting to improve the next-period correlations (reproducibilities).

3.3 Data Structure

Consider a set of eligible past spans $\mathbb{S}_-$. For a given span $S(t, -k) \in \mathbb{S}_-$, its corresponding future span is $S(t, +k)$. These give metrics $X(S(t, -k))$ and $X(S(t, -k))$. Further, we can collate these metrics over multiple meridians:

$$X(-k) = \texttt{row append}(\{X(S(t, -k)) : t = 0, 1, \ldots\};$$

$$X(+k) = \texttt{row append}(\{X(S(t, +k)) : t = 0, 1, \ldots\}.$$

Provided the rows (quarterbacks) of $X(-k)$ align with those of $X(+k)$, these matrices are sufficient to calculate correlations between spans of $-k$ and $+k$.

We can include multiple spans by an analogous column append operation:

$$X_{-H} = \texttt{column join}(\{X(-k) : k = 1, 2, \ldots, H\};$$

$$X_{+H} = \texttt{column join}(\{X(+k) : k = 1, 2, \ldots, H\}.$$

We constrain the rows (quarterbacks) to line up. H, the index of the largest span, we call the *horizon.* By considering jointly multiple spans, we can assess the incremental value of longer spans (more data aggregated) over shorter spans (more current, more recently measured, fresher data).

3.4 Objective Function

For a given span k and non-negative weights w such that $w^T 1 = 1$, we calculate the past and future composites by the matrix products $z(-k) \equiv X(-k)w$ and $z(+k) \equiv X(+k)w$, respectively. For a given horizon H, we wish to consolidate these z-composites of various spans with the non-negative weights u such that $u^T 1 = 1$:

$$z_{-H} \equiv \sum_{k=1}^{H} z(-k)u_k \text{ and } z_{+H} \equiv \sum_{k=1}^{H} z(+k)u_k.$$

For any vector w weighting metrics and vector u weighting spans, we obtain z_{-H} and z_{+H}, and their correlation. These correlations are the higher-is-better reproducibilities of the respective (w, u)-composite.

This can be formalized as follows: Define

$$z_{-}(w, u) = \sum_{k=1}^{H} \sum_{j} u_k X(-k)[, j]w_j \text{ and } z_{+}(w, u) = \sum_{k=1}^{H} \sum_{j} u_k X(+k)[, j]w_j$$

.

We wish to maximize

$$\mathbb{COR}(z_{-}(w, u), z_{+}(w, u)) \text{ subject to } w \geq 0 \text{ and } u \geq 0 \text{ and } w^T 1 = u^T 1 = 1.$$

The requirement that the weights w and u sum to 1 is not strictly necessary but ensures uniqueness. The requirement that these weights be non-negative ensures that all candidate composites are appropriately monotone in each metric component. Finally, the value that this objective function achieves is the reproducibility of the (w, u)-tuple.

This algorithm strongly resembles canonical correlation (Anderson, 2003, Chapter 12). One distinction is that the above program is *symmetric* between past and future; the same weights w are applied to past metrics and future metrics and the same weights u are applied to composites of past spans and composites of future spans. A second distinction from canonical correlation is the requirement that $w, u \geq 0$. The above algorithm we call *Spokane,* for *s*ymmetric, *po*sitive, *can*onical correlation.

4 Example

We now return to the five-season NFL quarterback data, 2012-2013 through 2016-2017. Table 2 gives the results for horizons from 1 to 8 weeks, the latter marking the season halfway point. The metric weights w are presented in Table 2(a), those of the spans u in Table 2(b). These weights are based on the appropriate correlation matrices, so the ratios COMP, YARD, TD, and INT have all been standardized by dividing by their respective standard deviations. All correlations are weighted, using Mantel-Haenszel weights, $[1/ATT(-k) + 1/ATT(+k)]^{-1}$, as appropriate.

In Table 2(a), we see three patterns. First, the weights on YARD are all zero; this implies that including any positive credut on yardage reduces the reproducibility. Second, the weights on INT fall off to zero once the horizon $H \geq 3$; this likewise suggests that compiling interception rates across three or more games does not contribute to predictions of future quarterback performance. Third, weights on completions (COMP) grow as span increases, and those on touchdowns (TD) slide down; for the intermediate spans of three to five games, their respective weights are roughly equal. This implies that completion rates (COMP) are the best simple summary of quarterback performance over large numbers of games, while a 50:50 mix of completions and touchdowns offers a plausible simplified composite.

Table 2(b) demonstrates a simple pattern: For predicting future performance, aggregating over a larger number of games helps the most. (For the horizon $H = 8$, there is a small amount of nuance on this point, where there spans of 6 and 7 are slightly up-weighted to 11 and 13 percent.)

For any given horizon, we compute also the reproducibilities, presented in Figure 2 and Table 3. Of the base metrics, COMP has consistently better reproducibility than the composite QBR. That said, an equal weighting of COMP and TD ("50:50") has reproducibility almost as good, and better than QBR. By construction, the Spokane solution, based on the weights of Table 2, has the best reproducibility. At horizon 8, QBR achieves a reproducibility of 0.4; Spokane comes in above 0.5. COMP and 50:50 give reproducibilities in between, about 0.45.

horizon	COMP	YARD	TD	INT	span=1	=2	=3	=4	=5	=6	=7	=8
1	0.36	0.00	0.23	-0.40	1.00							
2	0.49	0.00	0.08	-0.44	0.02	0.98						
3	0.51	0.00	0.49	0.00	0.00	0.00	1.00					
4	0.43	0.00	0.57	0.00	0.00	0.00	0.00	1.00				
5	0.44	0.00	0.56	0.00	0.00	0.00	0.00	0.10	0.90			
6	0.74	0.00	0.26	0.00	0.01	0.00	0.00	0.00	0.00	0.99		
7	0.85	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	
8	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.13	0.76

Table 2: (a) Metric weights w and (b) span weights u.

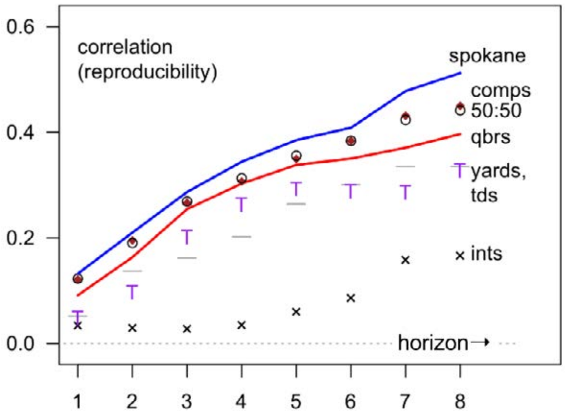

Figure 2: Reproducibilities of 7 composites.

horizon	QBR	COMP	50:50	Spokane
1	0.091	0.121	0.123	0.132
2	0.163	0.195	0.190	0.210
3	0.254	0.266	0.269	0.287
4	0.303	0.307	0.313	0.344
5	0.338	0.349	0.356	0.385
6	0.350	0.384	0.384	0.408
7	0.371	0.431	0.423	0.478
8	0.396	0.450	0.441	0.512

Table 3: Reproducibilities of the top 4 composites. 50:50 weights COMPs and TDs equally.

5 Remarks

gamuts: Measurements have purposes. SATs assess students to guide college acceptance. Quarterbacks are rated to guide game assignment and play selection. Our particular interest here is *make-a-gamut* problem. Gamuts are never an end unto themselves; rather they help us stratify data and refine models.

Previous efforts (Heavlin, 2016) make gamuts from a predictor-response model. The present work considers gamut making without such a response. Note that whereas a predictor-response formulation implicitly includes some notion of minimizing prediction error, the no-response case requires some extra work in order to inject prediction error into the formulation. We have found it convenient to recast the lower-is-better prediction error into the higher-is-better criterion of *reproducibility*, i.e. the correlation between the current prediction and the actual future value.

Methods: The standard method for constructing a latent variable from multiple measurements uses principal components analysis (PCA), taking as weights the first eigenvector of a correlation (or covariance) matrix, solving $(R - \lambda I)x = 0$ for the largest scalar λ and vector x. Invariant to changes of units, this method gives priority to maximizing the variation of the proposed scores. This same data structure, essentially the correlation matrix, makes no use of background or replication error. Diamantaras and Kung's (1996) oriented principal component analysis (OPCA) solves the eigensystem $(R - \lambda R_0)x = 0$, where R_0 represents the background noise. OPCA nudges solutions away from the eigen-directions with higher noise.

In our particular case, we take measurements at two points in time, before the meridian and after. This points us toward an explicit representation of the correlations between before and after — R_{-+} in section 3.2. Classically, the method of canonical correlations is the eigen-based generalization of PCA that keys on such cross-correlations; the underlying numerical method is the singular value decomposition (SVD).

As noted in section 3.4, Spokane differs from canonical correlation only modestly. (1) Spokane is symmetric, constraining the right- and left-hand eigenvectors to equal one another. (2) Spokane constrains the signs on the weights to a priori values. Both (1) and (2) ensure the ultimate *meaningfulness* of its solution. The corresponding objective function maximizes the solution's *reproducibility.*

Choosing the criterion of reproducibility reflects both (a) a mathematical nuance and (b) a philosophical value. (a) The nuance stems from defining reproducibility not as a measure of variation but rather as a correlation. Among other things, this means that constants, which have zero variance, have reproducibility zero and are therefore implicitly excluded. Further, the reproducibility as correlation tells us how well a past measurement predicts a future level. In this sense, the reproducibility concept is recast to emphasize prediction.

(b) Philosophically, measurements always look backwards in time — they require data — but their value derives from implying something about the future. In the case of students, IQ purports to measure the innate ability of students to grapple cognitive challenges: its value comes from its forecast of future performance. Likewise, quarterback ratings are available on fantasy football websites because they suggest which quarterbacks will play better in the next fantasy football match-up. These remarks seem obvious enough, but the conclusion is not recognizable: principal components and its public version, the quarterback rating, are not constructed to value reproducibility.

The reproducibilities reported in Table 3 are on the low side, about 0.4. In the context of reviews of submissions to a prestigious computer science conference, Tomkins et al.(2017) also report inter-reviewer correlations of 0.4. Alas, one consequence of measuring reproducibility is that one comes to confront uncomfortably low reproducibilities. On one hand, a certain humbleness about measuring and judging humans ensues; on the other hand, a resilient optimism: the poor performance on a given day predicts surprisingly little about the next.

QBR Results: As Table 3 indicates, most of our progress comes from increasing spans and sample sizes, and only secondarily from re-weighting the component metrics. Nonetheless, the maximum-reproducibility paradigm suggests simplifications: drop the yardage component, and perhaps interceptions. The completion rate COMP alone has better reproducibility than QBR.

These simplifications make sense: Arguably, quarterbacks have the most control over completion rates, the most reproducible part of their performance. In contrast, yardage has much to do with the receivers' and defenders' skills post reception. Both interceptions and touchdowns are relatively rare events, so can inject considerable noise. If simplicity is given some preference, pass completion rates (COMP) are to be recommended. If a composite is desired, the 50:50 composite, which assigns equal weights to completions and touchdowns, offers reproducibility almost as good.

6 References

Anderson, T W (2003). *An Introduction to Multivariate Statistical Analysis,* Chapter 12, Canonical Correlations and Canonical Variables. Wiley, New York.

Fantasy Football Today, (2017). Quarterback Stats: 2016 Regular Season, www.fftoday.com/stats/playerstats.php, accessed January 15, 2017.

Gardner, H (1983). *Frames of Mind: The Theory of Multiple Intelligences*, Basic Books, New York.

Heavlin, W D (2016). Modeling with gamuts, *Proceedings of the American Statistical Association*, American Statistical Association, Alexandria, VA, pp. 1125-1134.

Hotelling, H (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441, and 498–520.

Diamantaras, K T and Kung, S Y (1996). *Principal Component Neural Networks: Theory and Applications*, Wiley, New York.

Spearman, C (1904). "General Intelligence," Objectively Measured and Determined, *The American Journal of Psychology*, 15, 2, April, pp. 201-292.

Tomkins, A, Zhang, M, and Heavlin, W D (2017). Single versus Double Blind Reviewing at WSDM 2017, arxiv.org/pdf/1702.00502.pdf, accessed October 1, 2017.