BENEFIT-OF-THE-DOUBT APPROACHES FOR CALCULATING A COMPOSITE MEASURE OF QUALITY

By Michael Shwartz, James F. Burgess, Jr. (Presenting), and Dan Berlowitz

Funded by VA Health Services Research and Development grant IIR 06-260

COLMR
CENTER FOR ORGANIZATION, LEADERSHIP & MANAGEMENT RESEARCH

Context and Background I

- Standard Approaches for Creation of Composite Measures of Quality (Quality Indicators -- QIs)
  - Equal Weighting
  - Prevalence Based Weights
  - Judgment Based Weights
- Concept of Benefit-of-the-Doubt Approaches
  - Relative Performance represents a Measure of Revealed Preferences by the Organizational Unit on Relative Importance

Context and Background II

- Distinguish two types of Composite Measures
  - Reflective Measures (manifestations of construct)
  - Formative Measures (defined by individual QIs)
- Illustrate some approaches to create Formative Measures from QIs
- QIs are not Highly Correlated and Explicitly are Added to Include More QI Dimensions

Benefit-of-the-Doubt Measures

- Nardo et al. (OECD-2005) Review of Methods
- Benefit-of-the-Doubt Approaches Recognize Revealed Preferences w/Higher Weights
- Cherchye et al. (JORS-2007) and Semple (EJOR-1996) note this is the Natural Outcome of Nash evaluation game: Regulator v. Org.
- Mostly used to date to Compare Countries (e.g. Lovell (IJPE-1995), Despotis (JORS-2005))

Criticism and Intuition

- If Weights are Organization-Specific are Comparisons Across Units Possible?
  - Dropping Lowest Grade Example
  - Data Envelopment Analysis (DEA) does this
- If Final Comparisons are made on Relative Basis then what happens in practice?
  - No one knows in advance who benefits most
  - Actual rankings may not change much
  - Dropping or downweighting lower scores may buy good will from the organization/student at low cost

Purpose Statement

- Imagine we have a fixed set of QIs with a reporting period just ended
- Goals for the Regulator might be:
  - Facilitating consumer choice with gestalt value
  - Pay-for-Performance to reward high performers
  - Quality improvement learning to spread value
- Comparative Approaches
  - DEA (here all QIs are reported on the same scale)
  - Simple LP Optimizing subject to constraint that weights sum to 1 (needs QIs on the same scale)

Example: VA Nursing Homes (1998)

- 35 Nursing Homes in VA (Berlowitz et al. 2003)
- Five QIs Reflecting Patient Change Over Time
  - Pressure Ulcer Development
  - Functional Decline
  - Behavioral Decline
  - Mortality
  - Preventable Hospitalization
- All QIs are Risk Adjusted w/Published Models
- 32 Nursing Homes with no Missing Data used

Calculating the QIs

- Many ways can be used to calculate a QI, not of importance in this example
- Model generates Predicted Probability of 6 month adverse event given initial risk
- Add up observed adverse events (O)
- Add up predicted probabilities (E)
- We create O/E Ratios which are widely used

Comparisons of Composites

- Equal Weights Model
- Facility-Specific Prevalence Weights Model
- Overall Prevalence-Based Weights Model
- Simple LP Model (weights sum to 1)
- Weight Constrained DEA Model
  - Employ Rachel Allen/Thanassoulis Constrained Ratio of the Weights Measure
  - This does not permit some QIs to drop weights to near zero (the student drop the lowest grade model)

Table 2: Composite scores and facility ranks for high and low ranked facilities

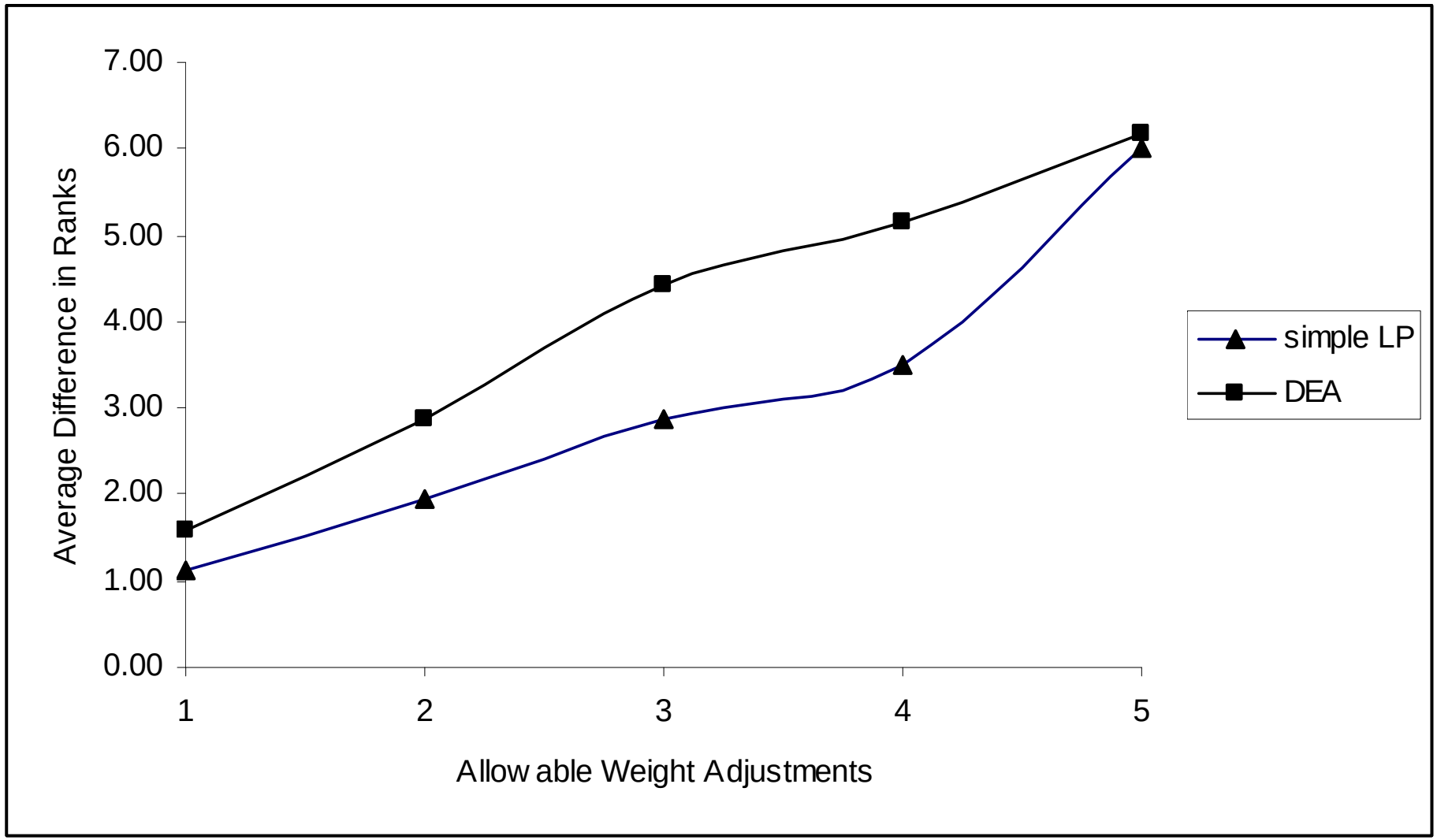
	Composite Score					Facility Ranks				
facility	overall prevalence-based weights	facility-specific prevalence-based weights	equal weights	simple LP model*	DEA*	overall prevalence-based weights	facility-specific prevalence-based weights	equal weights	simple LP model*	DEA*
10	0.576	0.495	0.509	0.351	1.000	1	1	3	1	3
6	0.654	0.530	0.495	0.451	1.000	2	2	2	3	6
28	0.662	0.875	0.445	0.412	1.000	3	7	1	2	1
19	0.754	0.790	0.755	0.596	1.000	4	4	8	6	5
8	0.762	0.957	0.599	0.479	1.000	5	15	4	4	2
24	0.779	0.848	0.621	0.636	0.969	6	6	5	7	9
11	0.805	0.782	0.917	0.680	1.000	7	3	16	8	4
17	0.857	0.846	0.728	0.566	0.990	9	5	6	5	8
4	0.997	1.002	0.914	0.868	0.900	21	20	14	22	28
15	1.035	1.037	1.190	0.936	0.939	23	23	27	27	18
16	1.047	1.056	1.091	0.951	0.895	24	24	23	28	29
13	1.086	1.086	1.198	1.001	0.903	26	26	28	29	27
31	1.118	1.134	0.958	1.023	0.881	27	27	18	30	31
30	1.150	1.402	1.236	0.855	0.952	28	30	29	19	14
32	1.158	1.323	1.162	0.858	0.930	29	29	25	20	23
18	1.272	1.303	1.321	1.187	0.837	30	28	30	32	32
20	1.385	1.556	1.383	1.107	0.889	31	32	31	31	30
2	1.443	1.520	1.717	0.923	0.949	32	31	32	25	16

*: results for the benefit-of-the-doubt approaches are for allowable weight adjustments of $\pm$ 0.75 of overall prevalence-based weights

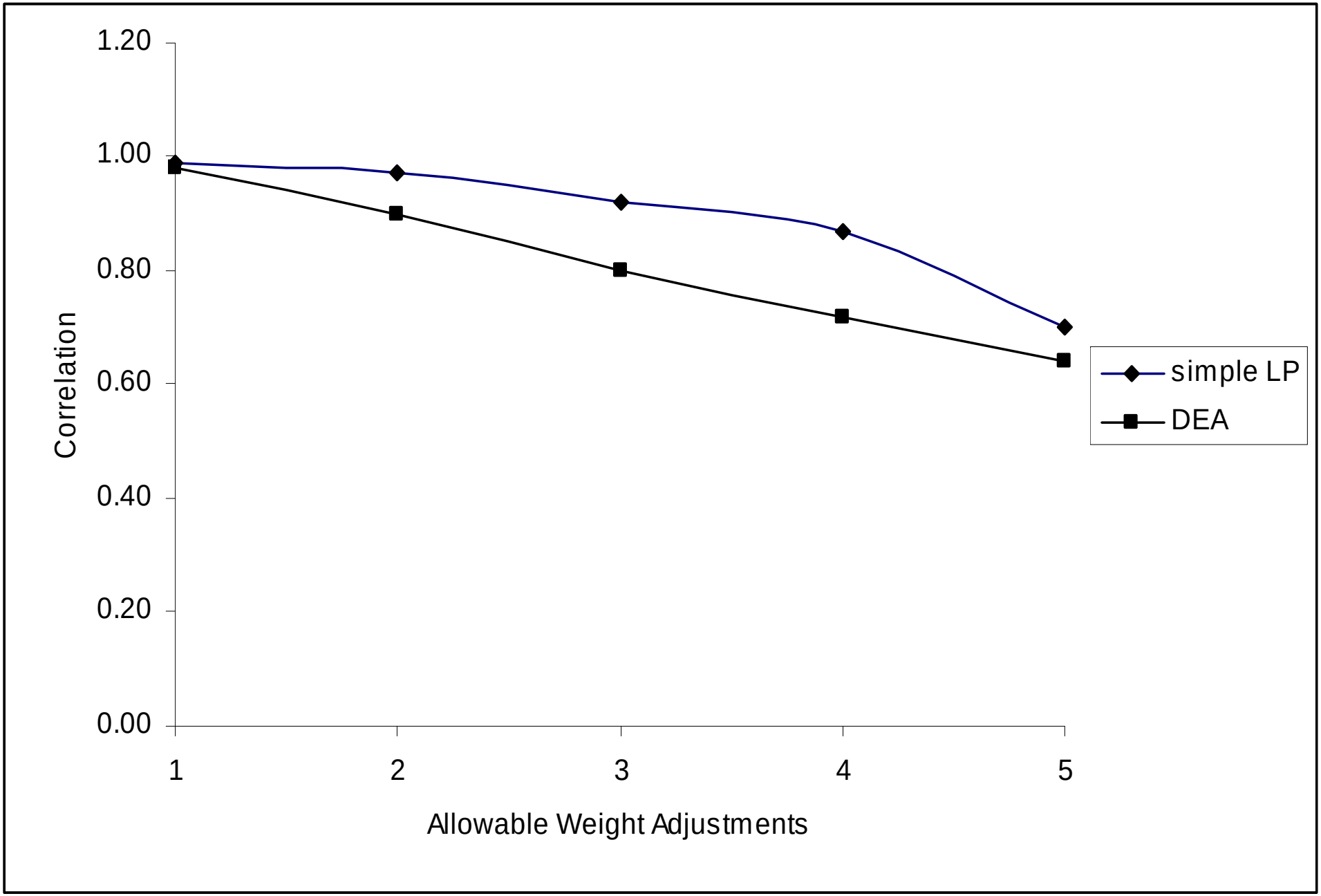
Weight Constrained Models Tested

- We Test Differences in Ranks/Correlation
- Levels: Allowable Weight Adjustments
- 1: ± 0.25*overall prevalence-based weights
- 2: ± 0.50* overall prevalence-based weights
- 3: ± 0.75*overall prevalence-based weights
- 4: ± 0.90*overall prevalence-based weights
- 5: no constraints

Figure 1: Comparison of ranks using overall prevalence-based weights to ranks using each of the benefit-of-the-doubt approaches with different amounts of allowable weight adjustments (previous slide)
Part A: Average difference in ranks



Part B: Correlation



Outcomes of Benefit-of-the-Doubt

- There is no gold standard for weighting
- But “equal weighting” is a choice and may generate: “these weights do not reflect what is important to our patients”
- Face validity? A moving concept?
- Post Hoc Discussion of Weights can only be Self-Serving
- But if true preferences are reflected in performance this approach should lessen tensions and improve trust and engagement
- No Need to Blame the Messenger!

Other Outcomes and Benefits

- We know Risk Adjustment is imperfect, so some adjustment is made
- Using Weight Constraints in DEA Allows Policymakers to Choose how far to go
  - We used simple constraints but others possible
- DEA has been used before and has favorable properties (Nash outcome, flexible to scores)
- DEA also has negatives (best with large amounts of data to set benchmarks)
- Simple LP must be normalized but may be more transparent than DEA – Simplicity a Virtue

Incentive Effects and Gaming

- If a organization performs similarly on all QIs it gains no value from the approach
  - Unless scoring “high” relative performance suffers
- Managers will focus on QIs where they can improve & which are most important to them
- P4P Programs now leaning toward rewards for attainment and improvement (to balance incentives), this method can combine or use regulator weights between them for totals

Limitations and Improvements

- Simple O/E ratios can be improved upon
  - (O-E)/variance (O) or z-scores
  - Hierarchical modeling results (Bayesian or not)
- More data, more measures, more recent data, more data over time all can be incorporated
- CMS Nursing Home Compare has a relatively complex algorithm while CMS Hospital Compare currently using simpler methods
  - Concept of “Five Star” systems

Final Thoughts

- Explosion of Quality Measures (QIs) in recent years
- Measurement of Composites is going to continue to be debated
- Inherent limitations (safety net facilities, incomplete risk adjustment) support flexibility to generate trust and buy-in
- Benefit-of-the-Doubt Measures should be part of the discussion