THE NASH & CIBINIC REPORT
government contract analysis and advice monthly from professors ralph c. nash and john cibinic

Author: Ralph C. Nash, Professor Emeritus of Law, The George Washington University
Contributing Authors: Vernon J. Edwards and James F. Nagle

DECEMBER 2022 | VOLUME 36 | ISSUE 12

¶ 68 NUMERICAL SCORING IN SOURCE SELECTION: Lessons To Be Learned

Vernon J. Edwards

Addendum by Ralph C. Nash

Numerical ("point") scoring/rating systems have long been used as a proposal evaluation technique. Once upon a time the use of such systems in the source selection process was common. But after a number of bid protest decisions in the 1970s and 1980s involving numerical scores, some agencies began to restrict or even prohibit their use and mandated the use of adjectival and color-rating schemes instead of numbers.

Numerical scoring/rating schemes are commonly used in formal systems of multiple-criteria and multiple-attribute decisionmaking and are described in texts such as Von Winterfeldt and Edwards, Decision Analysis and Behavioral Research (1986), and Goodwin and Wright, Decision Analysis for Management Judgment (5th ed. 2014). Von Winterfeldt and Edwards explain the rationale for numerical scoring (rating) on page 20 of their book under the heading *Numerical Subjectivity*, as follows:

> The fundamental principle might be called numerical subjectivity, the idea that subjective judgments are often most useful if expressed as numbers. For reasons we do not fully understand, numerical subjectivity can produce considerable discomfort and resistance among those not used to it. We suspect this is because people are taught in school that numbers are precise, know from experience that judgments are rarely precise, and so hesitate to express judgments in a way that carries an aura of spurious precision. Judgments indeed are seldom precise—but the precision of numbers is illusory. Almost all numbers that describe the physical world, as well as those that describe judgments, are imprecise to some degree. When it is important to do so, one can describe the extent of that imprecision by using more numbers. Very often, quite imprecise numbers can lead to firm and unequivocal conclusions. The advantage of numerical subjectivity is simply that expressing judgments in numerical form makes it easy to use arithmetical tools to aggregate them. The aggregation of various kinds of judgments is the essential step in every meaningful decision.

You can read an extensive discussion of source selection proposal scoring/rating methods and an explanation for the switch to adjectives and colors in a 1982 law review article by James C. Babin,

Federal Source Selection Procedures in Competitive Negotiated Acquisitions, 23 A.F. L. Rev. 318. An excerpt:

> Numerical formulae are firmly embedded agency evaluation techniques, even though they are not required. Yet, subjective rating systems have been the subject of far less protests. Perhaps this fact should provide at least food for thought among the agencies.
>
> * * *
>
> *Numerical Formulae*. With regard to numerical systems, the Comptroller General has recognized their relative usefulness, stating that such ratings are useful guides in the evaluation process but are not conclusive as to the actual adequacy of individual proposals. They are used in an "attempt to quantify what is essentially a subjective judgment" and "only reflect the disparate judgments of the evaluators and thus a difference in scores may not reflect an actual difference in merit." Numerical point ratings are sometimes used to score only initial offers, while best and final offers are subjectively analyzed, or both initial and best and final offers may be numerically scored. The scores assigned by evaluators are most often used in proposal evaluation, but in certain instances, such scoring is "normalized" so that the highest rated proposal is equated to a maximum score. Again, agency discretion predominates, as systems can be broad or very detailed.
>
> * * *
>
> [I]t might be mentioned that AFR 70-15 [the Air Force's then-famous source selection regulation] is presently undergoing revision. No estimate on a publication date for the new regulation is possible. However, although no basic changes in policy or procedures are contemplated, it is likely the Air Force will encourage the use of color-coding techniques rather than numerical scoring, at least in major systems acquisitions. Seemingly, the feeling is that the use of color coding provides a greater margin for discretion and avoids the inelastic connotations of numerical scores. Thus, subjective evaluations of the strengths, weaknesses, and risks of proposals falling within a particular color code presumably may be accomplished with more flexibility. [Footnotes omitted.]

Several agencies now require the use of adjectival or adjectival/color scoring/rating methods, but numerical methods are still in use.

Federal Acquisition Regulation 15.305(a) permits agencies to use "any" rating method, and the GAO has never objected in principle to the use of numerical scoring. We believe that it was in *Reaves, Pogue, Neal and Rose*, Comp. Gen. Dec. B-176763, 52 Comp. Gen. 686, 1973 CPD ¶ 26, 1973 WL 8606, that the GAO first stated its familiar mantra: "We believe that technical point ratings are useful as guides for intelligent decision-making in the procurement process…." Today, however, the GAO applies that mantra to all scoring/rating systems, not just numerical systems. See *KPMG LLP*, Comp. Gen. Dec. B-420949, 2022 WL 16921986 (Nov. 7, 2022): "Ratings, whether numerical, color, or adjectival are merely guides for intelligent decisionmaking." However, some agencies construct peculiar schemes. Consider, for example, the scheme described in *R&K Enterprise Solutions, Inc.*, Comp. Gen. Dec. B-419919.6, 2022 CPD ¶ 237, 2022 WL 4378258, 64 GC ¶ 299.

Numerical Rating In A Task Order "Fair Opportunity" Competition

On May 21, 2021, the Air Force Air Combat Command issued a "Fair Opportunity Proposal Request" (FOPR) for an order to be issued under the General Services Administration's OASIS contract. The requirement was for "information dominance support"—a "suite" of training, operations, and administrative services. The prospective task order was to provide for the issuance of firm-fixed-price, level of effort, and cost-reimbursement tasks and have an ordering period of one year with four one-year extension options. Presumably, the acquisition was conducted pursuant to the "fair opportunity" rules in FAR 16.505(b), not the source selection rules in FAR Part 15.

The fair opportunity was conducted as an essay-writing competition. The GAO quoted the Air Force's FOPR as stating that offerors would be evaluated based on whether:

> The offeror's proposal demonstrates a holistic understanding of the three major mission areas represented under [Performance Work Statement] Paragraph 2…as they relate to the employment and management of cyber weapon systems under this effort. For each mission area, the offeror's identified mission complexity/challenge and proposed strategy to address it demonstrates their clear understanding and ability to support the missions aligned under this requirement.

So, as described by the GAO, the proposal was not to be a description of the service that the Government would receive. It was to be a paper "demonstration" of the offerors' knowledge and ability.

The Air Force received nine proposals, eliminated six as ineligible, and then evaluated the remaining three. The protest focused on the evaluations of two of those three offerors: R&K Enterprise Solutions (R&K), the protester, and Cyber Engineering and Technical Alliance, LLC, (CETA), the company selected to receive the task order.

The Method Of Proposal Evaluation

The FOPR said the agency would use a "best value tradeoff" process to select the contractor. The GAO described the agency's evaluation factors, proposal rating scheme, and evaluation process as follows:

> The FOPR stated that award would be made on a best-value tradeoff basis considering three factors: pass/fail, price, and technical. The award decision would be based on an "integrated assessment" of these three factors, with the technical factor significantly more important than price. [Footnote omitted.]
>
> The technical evaluation factor was divided into two subfactors, each of which were further divided into three criteria. The first subfactor was technical experience, approach, and mission understanding. Its three component criteria were technical experience/past performance, mission and scope understanding (major mission areas), and mission and scope understanding (government headquarters (HQ)-level support). The second subfactor was management and staffing plan; its three criteria were titled recruit, retain, and manage.
>
> * * *
>
> In evaluating proposals under the technical factor, the agency intended to assign proposals an overall numerical score, referred to as a weighted total evaluation score (WTES). First, the agency would assess each proposal a score of five, four, three, or zero on each criterion. These scores were defined as follows:

Score	Description
5 points	Proposal clearly exceeds the minimum requirements of the solicitation criteria and indicates an exceptional understanding/approach to meet mission and PWS requirements. The proposal contains multiple strengths and no weaknesses or deficiencies.
4 points	Proposal clearly meets the minimum requirements of the solicitation criteria and indicates a thorough understanding/approach to meet mission and PWS requirements. The proposal contains strength(s) that outweigh weaknesses and no deficiencies.
3 points	Proposal clearly meets the minimum requirements of the solicitation criteria and indicates an adequate understanding/approach to meet mission and PWS requirements. The proposal contains no strengths or weaknesses, OR the strengths do not outweigh the weaknesses and no deficiencies.
0 points	Proposal does not clearly meet the minimum requirements of the solicitation criteria and has not demonstrated an adequate understanding/approach to meet mission and PWS requirements. Proposal contains one or more deficiencies.

> After scoring proposals on each of the criteria, the agency would follow a mathematical formula set forth in the solicitation to calculate an overall WTES. Next the agency would evaluate each offeror's total price. The FOPR provided that the agency would then make its tradeoff decision by "conduct[ing] an integrated assessment [of] the WTES and [the total evaluated price] to determine which proposal represents the best value to the [g]overnment, where the technical rating of an offeror outweighs the price difference."

The protester challenged (1) the proposal evaluations, (2) the agency's decision not to conduct discussions, and (3) the agency's tradeoff decision. The GAO rejected the protester's first two complaints but sustained the protest against the tradeoff decision.

Does The Use Of Numbers Make A Rating Or Scoring Method Mathematical?

Read the agency's WTES score descriptions in the right hand column of the above table very carefully. The numerals 0, 3, 4, and 5 do not represent quantities. In that sense, they are not *numbers*; they are just category labels. Those labels could just as easily have been Poor, Fair, Good, and Excellent. (The GAO did not explain why the agency's scale goes from 0 to 3, omitting 1 and 2.)

The agency's scale is ordinal, not interval or ratio. (It was like the Mohs scale of mineral hardness. See the Wikipedia entry—https://en.wikipedia.org/wiki/Mohs_scale_of_mineral_hardness#:~:text=The%20Mohs%20scale%20of%20mineral,material%20to%20scratch%20softer%20material.) Thus, it is not truly mathematical. It may be used for rank ordering, but one cannot legitimately perform mathematical operations like addition, subtraction, multiplication, and division using numerals on such a scale. See Stevens, *On the Theory of Scales of Measurement*, SCIENCE, July 7, 1946, and Siddiqui et al., *Heuristics of Applying Statistical Tests Using Appropriate Measurement Scales* (2016), https://www.researchgate.net/publication/332781020_HEURISTICS_OF_APPLYING_STATISTICAL_TESTS_USING_APPROPRIATE_MEASUREMENT_SCALES.

Read, again, the agency's prescription for a rating of 4 points:

> Proposal clearly meets the minimum requirements of the solicitation criteria and indicates a thorough understanding/approach to meet mission and PWS requirements. The proposal contains strength(s) that outweigh weaknesses and no deficiencies.

The GAO decision does not state that the agency's FOPR defined strength or weakness. The FAR does not define "strength," but FAR 15.001 defines "weakness" as follows: "*Weakness* means a flaw in the proposal that increases the risk of unsuccessful contract performance. A 'significant weakness' in the proposal is a flaw that appreciably increases the risk of unsuccessful contract performance." As we have previously discussed, it is not necessarily the case that all strengths or all weaknesses are equal to one another. See *Postscript: Source Selection Decisions*, 32 NCRNL ¶ 26. Depending on how defined, strengths and weaknesses may not be true units. Thus, you cannot just count them and then compare offerors based on the counts. Two offerors are not necessarily equal because both have five strengths and no weaknesses. One might be better than the other because of the nature of its strengths. The same applies to weaknesses. Some weaknesses may be worse than others. Thus, it seems clear that two offerors could be given a score of 4, but that one might be better than the other. As Von Winterfeldt and Edwards point out, that difficulty can be addressed through the use of more numbers, e.g., decimal fractions, which would provide for scores such as 4.1, 4.2, and 4.3, etc. That would be similar to a technique sometimes applied to adjectival scoring: Good, Good+, Good++, Good+++.

According to the GAO, the Air Force aggregated offerors' WTES scores and gave R&K, the protester, 382.5 WTES "points" and CETA, the selectee, 443.33 "points," which meant that CETA had been more highly ranked than R&K and that, assuming the category assignments were properly made, CETA was almost certainly better overall than R&K. *But how much better? And how did the agency come up with those decimal fractions?* CETA's total evaluated price was $139,961,150. The protester's price was $112,053,150, a difference of $27,908,000. Time for tradeoff analysis. So, how much better was CETA than R&K? According to the GAO:

> The FOEB's [Fair Opportunity Evaluation Board's] report concluded with an award recommendation. The FOEB observed that "[t]he WTES difference between CETA and R&K is 14.7%. Therefore, the Government has the ability to gain 14.7% more value in technical superiority for 22.1% ($27M) more in price over the life of the contract by awarding to CETA." The FOEB stated that the technical factor was significantly more important than price, and concluded:
>
> > Although CETA's proposed pricing is approximately $27M more than R&K, CETA's technical proposal is clearly superior to R&K's technical proposal. The additional 60.83 points in technical superiority in awarding to the highest evaluated WTES of 443.33 from CETA outweighs the $27M price difference in awarding to R&K's lowest evaluated WTES of 382.5. In accordance with the [b]asis of [a]ward stated in the FOPR, CETA represents the best value for the [g]overnment.

Oops! The agency calculated percentages using numerals on an ordinal scale. Question: The 14.7% difference in CETA's and R&K's WTES scores represented what, specifically? What, specifically, would the agency receive in return for the $27,908,000 difference in price?

According to the GAO, the agency's selection official explained the selection of CETA as follows:

> CETA's WTES score of 443.33 is 10% higher than the next closest rated offeror. A price analysis was conducted using competitive pricing[,] historical pricing, and market research and it was determined that CETA's [total evaluated price] of $139,961,149.63 was fair, reasonable, balanced, and was not unrealistically low.
>
> For this reason, based on fair opportunity given to all offerors, it is my decision, as [d]ecision [a]uthority, to select CETA for award.

[Note: R&K, the protester, was not the "next closest rated offeror." The WTES score difference between CETA and R&K was 14.7%.]

We don't understand what the Air Force thought it was doing by calculating a percentage of nonprice difference based on a scale on which numerals were nothing more than categorical labels and on which two offerors could receive a score of 4, but one could be better than the other. A 10% advantage in WTES scores did not necessarily indicate a 10% advantage in nonprice value. And we do not understand why the agency thought it made sense to make quality-price tradeoffs by comparing percentages of technical differences that were calculated using numerical labels on an ordinal scale with percentage differences in prices, which are true numbers on a ratio scale. In saying this we are relying on the information in the GAO decision.

A Predictable Outcome

R&K protested, among other things, the Air Force's tradeoff procedure. The GAO denied all of R&K's complaints except that one, stating:

> In a best-value tradeoff procurement, it is the function of the source selection authority to perform a tradeoff between price and non-price factors, that is, to determine whether one proposal's superiority under the non-price factors is worth a higher price. *J.R. Conkey & Assocs., Inc. dba Solar Power Integrators*, B-406024.4, Aug. 22, 2012, 2012 CPD ¶ 241 at 9. Before an agency can select a higher-priced pro-

> posal that has been rated technically superior to a lower-priced but acceptable one, the award decision must be supported by a rational explanation of why the higher-rated proposal is, in fact, superior, and explaining why its technical superiority warrants paying a price premium. *Coastal Env'ts, Inc*., B-401889, Dec. 18, 2009, 2009 CPD ¶ 261 at 4.
>
> A source selection based on a mechanical application of point scores, without any qualitative assessment of proposals (i.e., without a consideration of the proposals' strengths or weaknesses), is unreasonable. *West Coast Gen. Corp*., B-411916.2, Dec. 14, 2015, 2015 CPD ¶ 392 at 12. Even if the source selection document contains summaries of the strengths and weaknesses of the proposals, our Office will sustain a protest where the record does not reflect a qualitative comparison of those strengths and weaknesses.
>
> * * *
>
> Accordingly, we sustain R&K's protest of the agency's tradeoff decision.

Based on decades of prior decisions, that outcome was entirely predictable.

"Those Who Do Not Learn History Are Condemned To Relive It"

Three things about this case make us shake our heads. *First*, in the 21st Century, what Contracting Officer and agency procurement attorney does not know that a source selection authority must not (1) make a tradeoff process selection decision without comparing substantive differences among offerors and their offers and documenting tradeoffs or (2) justify selection decisions based on comparisons of ratings? *Second*, what agency prepares a selection decision document that explains a selection decision in terms of ratings/price ratios? *Third*, what agency reviewing attorney finds a selection decision document to be legally sufficient that states the basis for the decision in terms of a rating/price ratio?

Forty-six years ago, in a famous bid protest decision, *Grey Advertising, Inc*., Comp. Gen. Dec. B-184825, 55 Comp. Gen. 1111, 76-1 CPD ¶ 325, 1976 WL 13172, the GAO stated:

> We have consistently stated that 'Technical point ratings are useful as guides for intelligent decisionmaking in the procurement process, but whether a given point spread between two competing proposals indicates the significant superiority of one proposal over another depends upon the facts and circumstances of each procurement and is primarily a matter with the discretion of the procuring agency.' 52 Comp. Gen. 646, 690 (1973); 52 Id. 738, 737 (1973); *ILC Dover*, B-182104, November 29, 1974, 74-2 CPD 301; *Tracor Jitco, Inc.*, 53 Comp. Gen. 896 (1975), 75-1 CPD 253; *Management Services, Incorporated*, 55 Comp. Gen. 715 (1976). As we said in T*racor Jitco, Inc., supra*:
>
> *** Uniformly, we have agreed with the exercise of the administrative discretion involved—in the absence of a clear showing that the exercised discretion was not rationally founded—as to whether a given point spread between competitive range offerors showed that the higher-scored proposal was technically superior, on a finding that technical superiority was shown by the point spread *and companying technical narrative*. We have upheld awards to concerns submitting superior proposals, although the awards were made at costs higher than those proposed in technically inferior proposals. [Emphasis added.]

See, too, almost 20 years later, *Pearl Properties,* Comp. Gen. Dec. B-253614.6, 94-1 CPD ¶ 357, 1994 WL 273228:

> [T]echnical point ratings are useful guides for intelligent decisionmaking, but too much reliance should not be placed on them; whether a given point spread between two competing proposals indicates a significant superiority of one proposal over another depends upon the facts and circumstances of each procurement…. Award should not be based on the difference in technical merit score alone, but should reflect the procuring agency's considered judgment of the significance of that difference. In other words, the selection official must determine what a difference in technical point scores might mean in terms of performance and what it would cost the government to take advantage of it.

Those principles were not new in 1976 or in 1994. They have been stated many times. See also the decisions of the U.S. Court of Federal Claims, such as *Wackenhut Services, Inc. v. U.S.*, 85 Fed. Cl. 273, 297 (2008):

> The [Source Evaluation Board's] point scores are entitled to deference, but only if the underlying decisions properly are explained in the Administrative Record. *See, e.g., Femme Comp, Inc. v. United States,* 83 Fed.Cl. 704, 768 (2008) (need for adequate agency documentation); *210 Earll, L.L. C. v. United States,* 77 Fed.Cl. 710, 720 (2006) (holding that an agency is required "to provide 'a coherent and reasonable explanation of its exercise of discretion [.]' ") (citation omitted); *Opti–Lite Optical,* 99–1 C.P.D. ¶ 61, 1999 WL 152145, at *3 (1999) ("While adjectival ratings and point scores are useful as guides to decision-making, they generally are not controlling, but rather must be supported by documentation of the relative differences between the proposals, their strengths, weaknesses and risks, and the *basis and reasons* for the…decision.") (emphasis added); *see also* Ralph C. Nash & John Cibinic, "Source Selection: A Variety of Agency Guidance," 3 No. 8 Nash & Cibinic Rep. ¶ 60 (August 1989) at 4 ("There is a slow trend toward conferring a significant amount of discretion on source selection officials. This is exhibited by the number of documents prohibiting numerical scoring of certain factors and the *requirement in most of them that evaluators prepare substantial narrative justification for the scores they give.")* (emphasis added).

Conclusion

How did the Air Force make the mistake that they did? The simple answer is that the Government has not seen fit to properly educate and train its acquisition personnel, and too many of those personnel have not made personal studies of their business. As far as we have been able to determine, the professional reading lists posted at the Defense Acquisition University website do not include a single text on multiple-criteria/multiple-attribute decisionmaking and tradeoff analysis. It does not include Decision Analysis for Management Judgment, by Goodwin and Wright, an excellent text. It does not include the more technical Decision Analysis and Behavioral Research, also cited above, the writing of which was sponsored in part by a U.S. Navy contract, N00014-79-C-0529, and to which the U.S. Government has a royalty-free license throughout the world in all copyrightable material contained therein.

To put it colloquially—*What's up with that?*

Coda—A Final Observation

The Air Force issued its FOPR on May 21, 2021, and took until May 21, 2022 to award a task order. It then opted to fight R&K's bid protest, which the GAO decided on September 12, 2022. Thus, the process of placing a task order against a Government-wide acquisition contract (GWAC) took, so far, 480 calendar days to complete. That's no way for a military outfit seeking "information dominance" to get there "firstest with the mostest."

Multiple award task order contracts, GWACs, and the "fair opportunity" process prescribed by FAR 16.505(b) were thought to be a great streamlining innovation of the mid-1990s, but in 1995, in *The New Rules for Multiple Award Task Order Contracting*, 9 N&CR ¶ 35, we warned:

> The multiple award preference policy states that every awardee must be given a "fair opportunity" to be considered for the award of each task order in excess of $2,500. The proposed rule leaves the choice of evaluation factors to the CO's discretion. The CO need not publish a synopsis, solicit written proposals, or conduct discussions with awardees prior to the award of a task order, proposed FAR 16.505(b)(1). The rule precludes protests against task order award decisions. Agencies must appoint task order "ombudsmen" to handle complaints from awardees about task order selections, proposed FAR 16.505(b)(4).

> Notwithstanding these liberal policies, it is not difficult to imagine Government procurement officials conducting a mini-source selection before the issuance of each task order. Some will almost certainly consider a more formal procedure to be necessary to ensure fairness. One can easily imagine requests for proposed task order "performance" plans or "management" plans, especially for task orders of significant dollar value. One can also imagine requests for extensive cost breakdowns, certified cost or pricing data, and proposal audits. If too complex and demanding, such procedures would significantly increase an agency's administrative costs, extend the lead time associated with task order issuance, and force awardees to incur significant costs in the preparation and negotiation of task order proposals.

Unhappily, we were right. Ignorance is not bliss, and innovation is not a substitute for competence. *VJE*

ADDENDUM

I can understand that the Contracting Officer and those above him or her in the contracting office might not know the fundamental rules of the competitive negotiation process. Vern is correct in describing the sorry state of some of our acquisition workforce. But the fact that no Air Force lawyer saw the use of this improper source selection technique or stopped this procurement in its tracks when the protest was filed is truly discouraging. Surely, one of the basic tenets for attorneys advising and representing a Government agency is to prevent the publication of a protest decision that tells the public that the agency is totally incompetent in carrying out its mission of procuring goods and services to meet what often are urgent needs. By this I don't mean that attorneys should head off inefficient procurement processes. As we have often said, most competitive procurements take too long and use up too many resources, and attorneys can't and shouldn't be required to change that. But when they see the agency violating a long-established rule, as in *R&K Enterprise*, they should go to the head of the agency if necessary to warn of the embarrassment that a published decision would inflict on the agency. That's what I call giving sound legal advice. *RCN*