Money for nothing: how quality-price trade-offs in bid scoring increase the risk of overpaying for public contracts

Janet Izatt. University of Warwick – Warwick Business School

Timothy L. Mullett. University of Warwick – Warwick Business School

Written: November 2024

Abstract

Public procurement is an astonishingly big business with governments globally spending trillions of dollars each year. The UK, which has been at the forefront of contracting out and outsourcing, spends nearly £400 billion annually. Concern that the Government is paying too little for contracts obscures the likelihood that it may also pay too much because of the compensatory price/non-price mechanism used to evaluate bids in competitive tenders. The Goldilocks problem of paying too little or too much for contracts is a conundrum for the UK's new Procurement Act (2023) which is underpinned by a commitment to achieving value for money.

The competitive tendering approach of using weights for price and non-price (quality) factors to signal a buyer's priorities fails to account for a nominal exchange rate between price and non-price scores which is established when scores for both elements are combined in a compensatory composite score. Small differences in non-price ratings which could arise from rater noise rather than genuine quality differences could lead to paying disproportionately more than necessary for a contract. Tendering procedures without negotiation lack a mechanism to resolve this problem.

By considering bid evaluation under different price/non-price weightings and contract values we show how contracting authorities could inadvertently pay too much. Implications and recommendations are considered.

Introduction

Public procurement has long faced criticism that competitive tendering favours lowest price over quality. When the collapse of the “too big to fail” government contractor Carillion in 2018 was reviewed, the Government was accused of paying too little leaving insufficient margins for contractors and driving down the quality of services (House of Commons Public Administration and Constitutional Affairs Committee, 2019). Carillion, which had liabilities of £5 billion pounds when it collapsed with the loss of thousands of jobs, was operating on profit margins of as low as one per cent on some Government contracts (National Audit Office, 2018).

Recent years have seen attempts to overcome those criticisms with changes in legislation downplaying the importance of price. The focus has changed over the years from contracts being awarded based on “lowest price” to “most economically advantageous tender” (MEAT), to “most advantageous tender” (Procurement Act, 2023, 19 (2)) with the objective to deliver “value for money” (Procurement Act, 2023. 12 (1)).

An accusation that contracting authorities may overpay for contracts is therefore novel and has significant implications. Savings from the government’s £393 billion annual spend on goods and services (National Audit Office, 2024) could fund additional services or upgrades to existing services. Overpayment in this paper means the extra amount a contracting authority (a public sector buyer) may agree to at contract award following a competitive tender without negotiation with no identifiable benefits over a cheaper bid.

The potential to pay too much or too little is caused by a bid evaluation procedure in which a bid’s objective price score and a subjective non-price score are combined in a compensatory composite score in which trade-offs occur between price and non-price scores. By indirectly changing the stated price and non-price weightings, this can change the competition’s award criteria. In tender procedures without negotiation, tenders are awarded to the bid with the highest score, rather than as a deliberate choice after reviewing all relevant information. Therefore, short of not awarding a contract to a bidder, contracting authorities may be unable to avoid overpaying.

The purpose of this paper is to describe how the circumstances arise under which contracting authorities may pay too much and demonstrate the size of potential overpayment under different weights and contract values.

This paper considers the implications of results from three other papers from one of the authors (Izatt)[1] which finds significant rating noise, or variability, in the subjective rating of bids (Izatt, 2024a, 2024b, 2024c). It adds to the literature on public procurement and policy by focusing on the effects of the often-overlooked aspects of subjective evaluation on tender outcomes. Extant research on tender evaluation has focused on weighting and optimising criteria. However, this focus ignores the sensitivity of competition outcomes – who wins a contract and at what price - to small differences in subjective non-price rating scores.

The rest of this paper is as follows. In the section *Public tender evaluation*, I briefly describe standard public procurement tender evaluation procedures. In *Literature*, I provide a short review of literature on subjective evaluation and rating aggregation. The effects of different price and non-price weights are demonstrated using illustrative examples in *Approach* to show their impact on contract award price. The results are reviewed in *Discussion* and implications, limitations and recommendations for how to restore balance between price and non-price factors are considered. The paper ends with a brief *Conclusion*.

Public tender evaluation

Public competitive tendering is the primary process by which public contracting authorities (Government buyers) award contracts for the goods and services they buy. In the UK, as in many countries, public procurement is regulated by statute and this has resulted in detailed rules and practice, including how tenders should be run and bids evaluated (Arrowsmith, 2005). In the UK, many of these rules evolved from practice (Arrowsmith, 2005). The aim of much of the regulation has been to eliminate bribery and corruption, increase transparency about how taxpayers' money is being spent (Arrowsmith, 2005) and, when the UK was a member of the EU, to ensure a level playing field for trade across the trading block.

The UK has been at the forefront of using competitive tendering in both the private and public sectors in the 1980s and 1990's due to the move to contracting out (Sturgess, 2017). Compulsory Competitive Tendering was introduced in the Local Government, Planning and Land Act (1980) to ensure competition between in-house providers (direct labour organisation) and outside providers (Arrowsmith, 2005). However, competitive tendering is now used globally in international agreements such as World Trade Organization's General Procurement Agreement (GPA) and EU Public Procurement Regulations (2015). The scale of the business-to-government market is enormous and growing. Annually, the UK Government spends £400 billion (National Audit Office, 2024), it is €2 trillion annually in the EU, and around $1.7 trillion annually through the GPA (World Trade Organization, 2024).

[1] Included in doctoral research Janet Izatt's PhD Thesis submitted November 2024.

In the UK, the procedures by which public authorities invite and evaluate tenders and award contracts for goods, works and services which exceed minimum threshold values are set out in the Procurement Act UK (2023), which comes into effect in February 2025[2]. The Act replaces the European Union (EU) Procurement Directives (2015) which applied to the UK during its membership of the EU. However, to meet its obligations under global agreements the process is similar to the previous regulations (Procurement Act, 2023, schedule 9).

There are different competitive tendering procedures available giving contracting authorities the option to restrict both the type and number of companies which can bid (open and restricted procedures), to include or exclude negotiation, to have a single or multiple stages, and in certain situations to award a contract without competition (direct award) (Procurement Act, 2023, ch 2, s 3). This paper will focus on those procedures (open or restricted) where a bid is either accepted or rejected without a negotiation stage.

Public contracting authorities are required to advertise opportunities on the government tender portal Contracts Finder (UK Government, 2024) setting out how bids will be assessed. Where there is more than one criterion, their relative importance should be indicated either by weighting, rank ordering or some other way (*Procurement Act*, 2023).

Contracting authorities decide how bids will be evaluated. However, Government Guidance (Government Commercial Function, 2021) and a review of the tender evaluation process described in a sample of tender documents published on Contracts Finder provide insight into common practice. Price and non-price elements are submitted and evaluated separately. Non-price elements can include a range of sub-categories such as experience, health and safety, technical, social value and commercial terms and conditions. The price and non-price scores are expressed according to the competition's price and non-price weights and then aggregated into a total score. Evaluators of non-price sections do not see bid prices, removing the influence of willingness to pay considerations from their ratings. My review of several tenders shows that 20:80, 30:70 and 40:60 are common price/non-price weights.

Further details of the price and non-price evaluation and aggregation are provided later in this paper in the *Approach* section where the effects of evaluation are investigated.

2 The Procurement Act applies to England and those areas where specific powers have not been devolved to Scotland (Scotland Act 1998), Northern Ireland (Northern Ireland Act 1998) and Wales (Government of Wales Act 2006).

Literature

Supplier selection has received considerable attention from academic researchers often under the umbrella term the “supplier selection problem”. Two key areas of research are the criteria, other than merely price, used to evaluate bids (Weber, Current and Benton, 1991; Watt, Kayis and Willey, 2010) and the analytical approaches and methods used for evaluation (for reviews see Weber, Current and Benton, 1991; Ho, Xu and Dey, 2010; Bruno et al., 2012). This body of research focuses largely on the design of the tender process and the scoring rules – identifying and weighting the overall criteria, defining the performance levels for each criterion and assigning values for each level.

A review of 78 papers on supplier selection published between 2000 and 2008 found that the most commonly studied multicriteria decision making tools for supplier selection were the formalised structured mathematical methods Data Envelopment Analysis (DEA) followed by Analytical Hierarchical Process (AHP) (Ho, Xu and Dey, 2010).

Data Envelopment Analysis (DEA) evaluates efficiency by analysing inputs and outputs. Originally used to evaluate and compare production processes, efficiency has been extended to quantify the performance of decision-making units (DMUs) such as, for example, organisations, divisions and activities (Niewerth, Vogt and Thewes, 2022). AHP is a mathematical process for optimising decision outcomes such as when maximisation (for example maximising profit) or minimisation (achieving lowest cost) is required (De Boer, Labro and Morlacchi, 2001). Both DEA and AHP have also been used in conjunction with other approaches such as AHP with fuzzy set theory (Bruno *et al.*, 2012) when it is necessary to evaluate qualitative criteria which have more uncertainty.

The problem of contracts being awarded to the lowest price bid is recognised in the procurement and supply chain literature and there have been efforts to develop decision support tools to improve decision outcomes in both public and private competitive tendering. Both DEA and AHP have been proposed as approaches to help overcome the possibility of awarding contracts to a low price, but potentially poor quality, bid in a public tender by removing weaker bids in an early stage of evaluation (Falagario et al., 2012; Cheaitou, Larbi and Al Housani, 2019; Dotoli, Epicoco and Falagario, 2020).

While powerful, mathematical approaches have their limitations. Both DEA and AHP are at their best when dealing with criteria which can be measured objectively (price, duration, weight), not when qualitative criteria (quality, reputation) which lack objective measures and therefore introduce uncertainty are rated subjectively. DEA measures efficiency, but this is not necessarily the same as supplier effectiveness (Ho, Xu and Dey, 2010).. When using

these approaches differences in supplier scores can be negligible resulting in ties, pair-wise comparison mechanism can be impractical and the use of aggregation methodologies can significantly reduce the variance in the weights assigned to the criteria (Bruno et al., 2012). Furthermore, AHP is complex and time consuming and not easily understood by procurement practitioners (Ho, Xu and Dey, 2010)

Use of mathematical approaches assume that evaluation criteria can be identified and agreed upon by multiple stakeholders within the buying organisation. Much, but certainly not all, of the research is in the manufacturing and construction industry rather than services, and examples are often set in private sector conditions which are not subject to public sector procurement restrictions. Based on their case study using AHP, Bruno et al (2012)suggested that that the supplier selection methodology should be dynamic with the result considered a starting point to be constantly monitored and improved. This has merit in the private sector environment where tenders may be used to inform and shape thinking but it would be problematic with public procurement where the evaluation scoring rules must be published a priori and adhered to.

A key problem for public tendering is where declared weights are wrongly interpreted as conveying the relative importance of the criteria but without considering the trade-off made. Other research has focused on the problems arising from the use of relative scoring when using a weighted sum model in which price and non-price scores are combined. The inclusion of a weak, although potentially cheapest bid, will affect the scoring and rank ordering of bids and may change which bid is likely to win (Schotanus *et al.*, 2022). Niewerth et al (2022) propose an extension to the use of DEA to remove tenders that do not stand a reasonable chance of success to avoid this problem while Mateus et al (2010) call for the trade-off between price and quality to be quantified when designing the scoring rules.

This body of work identfies on the issues and pitfalls involved in designing effective tendering evaluation scoring rules and the calculation of total bid scores. As such, it is concerned with the critical early stages of tendering and the end result when scores are combined. Common to many of the papers is an assumption that criterion can be objectively defined and measured. The difficulties of dealing with subjective rating of qualitative criterion are, when mentioned, acknowledged as a weakness but then are largely ignord. Uncertainty and ambiguity are unwelcome in these approaches.

To my knowledge there has been little to no investigation of the implications of subjective ratings for tender outcomes, nor of the potential to overpay. There are gaps in the procurement and supply chain literature about *how* bid evaluators rate qualitative criteria and to what extent those ratings affect tender outcomes given the issues with weights and trade-

offs between criteria. My research addresses these gaps and extends the literature on supplier selection by investigating the neglected issue of variance in bid ratings when different evaluators rate the same information. Rather than sidestepping the inconvenient problem of individual subjective rating inconsistency, I explore the conditions under which it occurs and to what extent and, in doing so, address the issues of transparency and fairness in public tender evaluation.

An assumption of rational behaviour is that individuals can consistently apply the same weighting and criteria to their evaluation. Decision research provides abundant evidence that this is not the case (Kahneman and Tversky, 1979) and that decision making is contextual and contingent on several factors (Beach and Mitchell, 1978).

In other words, people can be inconsistent in how they value options when the same choice is presented in a different form. This inconsistency can lead to bias and noise – systematic deviation and random scatter – in human judgement when the true answer is unknown or unknowable (Kahneman, Sibony and Sunstein, 2021).

Judgment, or a rating, is provided at the time it is asked for using a combination of the information provided, the manner and framing in which it is presented, and selective recall from memory (Payne, Bettman and Schkade, 1999). Different cognitive process are used to express preferences depending on how individuals are asked, and what they are asked to do – choose or value, using separate or joint evaluation – and this changes the outcome or stated preference. For example, when asked to choose between two bets a person may focus on the probability of a payout whereas when asked to price a bet (judgement) they may focus on the magnitude of the payout (Goldstein and Einhorn, 1987). Whether the judgement is made repeatedly (recurrent) or once only (singular) may also affect judgements (Kahneman, Sibony and Sunstein, 2021).

The objective mapping of subjective values also depends on the object's evaluability, or the extent to which a person has relevant reference information to gauge target values and map them onto an evaluation (Hsee and Zhang, 2010). The evaluation mode – separate or joint – plays a key role in determining what information is available.

In separate evaluation an individual sees and evaluates only one option at a time against some ideal or objective standard and cannot perform trade-off analysis which is used in comparative analysis. In the absence of a standard or an exemplar against which to judge the option, the first one evaluated, or some other concept, can provide an anchor for evaluating other options (Schkade and Johnson, 1989).

With joint evaluation the rater can move backwards and forwards comparing items, adjusting their ratings as they become more familiar with what is on offer. The basis for that

comparison can differ with evaluators sampling only some of the available information, comparing alternatives on key attributes or with their own prior experience.

Individuals use a range of heuristics, or mental shortcuts, to help them decide: comparing what they are rating to what they know or to other available options (Stewart, Chater and Brown, 2006; Vlaev *et al.*, 2011), evaluating those aspects that are easier to evaluate than others (Hsee, 2019) and focusing on particular information to make the task faster or easier (Gigerenzer and Goldstein, 1996).

When it comes to assigning a rating, individuals often use a restricted scale and struggle to be consistent with their ratings, let alone with others (Parducci, 1965; Barkaoui, 2011). Which rating a person uses will depend on how the item to be rated compares to items already compared. For example, someone used to high quality may be more critical awarding a low rating for an average bid, while someone used to poor quality is likely to award a high rating for the same bid (Parducci, 1965; Parducci and Wedell, 1986).

A separate study by this author (Izatt, 2024b) of the effects of evaluation mode and evaluability on bid ratings found rating variability increased significantly as requirements became more detailed.

A further issue is what happens when a subjective score is combined with an objectively derived score into a composite score.

Composite measures

Composite scores combine different indicators into one number or rating to make a complex phenomenon more understandable than looking at each of its parts. Providing one number such as a total bid score is easier to grasp than a number of individual scores for individual indicators.

Critics of composite scores argue that outcomes can be readily manipulated by altering the weights of individual elements if there is not a clear procedure justified to everyone (Grupp and Mogee, 2004; Grupp and Schubert, 2010).

Even if indicators or measures of performance can be agreed on, weights are problematic (Dawes, 1979). The first problem lies in the purpose of the weights. It can refer to the ‘explicit importance’ that is attributed to every criterion in a composite index, exhibiting its importance relative to the rest of the criteria or the “implicit” importance of the attributes, seen in the ‘trade-off’ between the pairs of criteria in an aggregation process. The first should result in non-compensatory aggregation, with the second leading to compensatory aggregation.

Problems arise when compensatory approaches are used when they should not be. For example, individuals believe the weights represent relative importance but are instead used in a compensatory manner when aggregated (see Greco et al., 2019 for a comprehensive review). Switching from a linear to geometric aggregation is seen as a partial solution (Greco et al., 2019).

Extreme caution is needed when drawing conclusions based on these measures (Greco et al., 2019). When it comes to aggregation, composites suffer from a trade-off between compensability and complexity or a loss of information. Weighting has been criticised as signaling only spurious precision or an order of precision which is not basically inherent in the data (Bobko et al., 2007).

Aggregation issues are common in the field of psychometrics where an individual's competence is measured by combining numerical sub-scores derived from instruments with different purposes. Converting human phenomena into numbers is not an exact process, but the problem is compounded when combining scores. Critics of this approach argue that assessors should interpret, not simply apply transformation to numbers (Hodges, 2013). Goldstein and Einhorn (1987) described this mapping of basic evaluations onto numerical scales as "mapping incommensurables" warning that its inappropriate application beyond being a useful heuristic can lead to anomalies and inconsistent behaviour.

Concerns arise when different measures are combined. For example, price meets the criteria for a ratio scale which can be used in a range of statistical calculations. The non-price element can comprise multiple questions, each of which is scored, usually on a Likert scale, with the scores then totalled and weighted. These are ordinal, not ratio, numbers. The price and non-price elements are then combined to provide an overall score.

In doing so, core principles of measurement theory would seem to be violated (Stevens, 1946). The combining and comparing of non-standard and different types of rating scales inappropriately obfuscates the advertised weighting of different elements of the tender. Hence, the true weighting of different elements may change in calculating an overall score. Although the transformations which occur within tender evaluation are common in psychological research studies (for example see Highhouse, 2001 or Highhouse, 2008) and appear to be accepted, statisticians warn that they should be done with caution, especially regarding the conclusions drawn from them (Stevens, 1946).

The extant literature on composite scores identifies the underpinning issues with the use of composite price and non-price scores in tender evaluation. This study builds on this literature by investigating the consequences of compensatory composite scores for public tender contract award outcomes.

Approach

This section describes the common tender practice for scoring price and non-price bid elements and combining the two. It is based on Government guidance on bid evaluation (Government Commercial Function, 2021) and a review of a sample of UK invitation to tender documents published on the tender portals Contracts Finder and TED Europa. After describing how price and non-price elements are scored the consequences of combining the scores are demonstrated.

A price:non-price weighting should be stated in the invitation to tender documents, and it is left to the contracting authority to choose the relative weights (*Procurement Act*, 2023). Government guidance states that weights should reflect the characteristics of the service. For example, an 80/20 quality/price weighting sends a clear signal to suppliers that quality is significantly more important than price (Government Commercial Function, 2021).

Quality evaluation

According to Government guidance, evaluators should assess bid non-price elements against a standard or benchmark set out in the tender documents using an ordinal Likert scale specified in the tender documents. Bids should be evaluated one at a time (separate evaluation) without comparison to other bids (joint evaluation). A review of tender documents shows that a five-point scale is used most, although three, four and seven-point scales are also used. Rating is analytical rather than holistic which means evaluators do not assign an overall score to a bid. Instead, each evaluator rates the answers to several questions, often within several sections. The evaluators' non price rating ratings go through various steps such as aggregating the ratings for each section, averaging them across evaluators, and expressing them according to the non-price weight for the competition.

A moderation stage may be used in addition to this first stage. In moderation evaluators discuss, review and modify ratings to reach an agreed rating for each bid. If a moderation stage is used, the average of the evaluators' ratings is replaced by the agreed moderation rating. Rating has been found to be harsher in moderation in another study by this author (Izatt, 2024c).

The number of questions and sections used in tenders varies considerably across tenders. Weights may be applied to specific questions and/or sections. My review of tenders across several sectors shows that these processes will involve some combination of an individual's ratings for each question and section being weighted, then totaled and expressed as a percentage of the overall non-price weight.

Therefore, non-price assessment follows the formula:

$$NPS = \overline{NPR} \times NPW$$

Where:

$NPR = \sum Rq \div (Rn \times Rm)$; and

- *NPS* is the bid non-price score
- *NPR* is the individual evaluator non-price rating
- *Rq* is the evaluator's rating of the bidder's response to an individual question
- *Rn* is the total number of questions to be answered
- *Rm* is the maximum score available to a bidder based on the Likert scale. (Most often 5)
- *NPW* is the weighting given to non-price factors in the overall tender evaluation

For example, in a non-price evaluation grid with 12 questions and a 5-point Likert scale where non-price carries an 80% weight a bidder receiving an average evaluator rating across, say, three evaluators of 35 points would achieve a non-price score of (35/60) X 80% = 47.

Price evaluation

With price evaluation there are three components to the overall evaluation:

- Weight – the proportion of the total bid evaluation score that is derived from the price element.
- Benchmark – the reference point for evaluation is either a target price (set by the buyer) or lowest offered price from the compliant bids received.
- Mechanism – The bidder's price is scored using a ratio scale - either linear or proportional - to the defined benchmark. This is explained below.

Buyers can reject a price for being "abnormally low" but, to my knowledge, there is little evidence of this happening. Before a bid can be rejected as abnormally low the contracting authority must give the bidder adequate time to demonstrate the bid price is not abnormally low (*Procurement Act*, 2023) and this can slow the tender process.

Price evaluation follows one of two forms: linear or proportional assessment.

Linear assessment

With linear lowest price benchmark evaluation (linear assessment), each submitted and compliant tender is evaluated against the lowest price offered and given a score equivalent to the inverse of the ratio of the offered price to the lowest price. This score is then multiplied by the weighting given to the price element of the overall tender evaluation to arrive at a weighted price score.

Hence linear assessment uses the following formula:

$$PS = \left(1 - \left((PBid - PLow) \div PLow\right)\right) \times PWgt$$

Where:

- *PS* is Price Score
- *PBid* is the bidders offered price
- *Plow* is the lowest compliant offered price
- *PWgt* is the weighting of price in overall tender evaluation

Table 1 sets out a hypothetical tender price evaluation with six bidders using linear lowest price benchmark in a competition where price carries a 40% weighting.

Table 1: Sample price evaluation using linear lowest

Bidder	**Bid Price**	**Differential to the lowest price which meets the mandatory pass criteria.** *(Expressed as a percentage)*	**Score** (100 – %Differential)	**Points (40% weighting)**
A	£100,000	0 – 0%	100	40
B	£120,000	£20,000 – 20%	80	32
C	£140,000	£40,000 – 40%	60	24
D	£150,000	£50,000 – 50%	50	20
E	£175,000	£75,000 – 75%	25	10
F	£200,000	£100,000 – 100%	0	0
G	£300,000	£200,000 – 200%	0	0

Using the example in Table 1, with linear assessment the lowest price gets a 100% score so, with a price weighting of 40%, a score of 40 points. Bid prices which are more than twice the lowest price score no points. This is the most punitive of the two price evaluation approaches. Due to commercial sensitivity contracting authorities are not required to publish all bid prices, only the winning price, so we do not have data to show how often bid prices will exceed twice the price of the lowest priced bid.

Proportional assessment

Where proportional lowest price benchmark scoring (proportional assessment) is adopted, bid prices are also compared to the lowest price but they are scored on an inverse proportional basis and expressed as a percentage of the lowest offered price meaning that higher price bids still receive some price points. The lowest price compliant bid score is, as with linear assessment, awarded 100% of the price attribution.

This is expressed in the formula

$$PS = (Plow \div Pbid) \times PWgt$$

Where:

- *PS* is Price Score
- *PBid* is the bidders offered price
- *Plow* is the lowest compliant offered price
- *PWgt* is the weighting of price in overall tender evaluation

Proportional price assessment is less punitive than linear price assessment. Using this evaluation approach, higher price bids such as F and G will score some points and stay in the contest even if they exceed twice the price of the cheapest bid (Table 2).

Table 2: Sample price evaluation using proportional low cost and 40% price weight

Bidder	Bid Price	Inverse ratio of offered price to lowest price	Score	Price is 40% of total score
A	£100,000	0 – 0%	100	40
B	£120,000	0.83	83	33.2
C	£140,000	.71	71	28.4
D	£150,000	.67	67	26.8
E	£175,000	.57	57	22.8
F	£200,000	.50	50	20
G	£300,000	.33	33	13.2

Combining price and non-price scores

The price and non-price scores are summed to reach a total score with the highest scoring compliant bid winning, subject to the contracting authority accepting the offer. As the tender is an invitation to treat under contract law (*Contracts (Applicable Law) Act*, 1990), the buyer is not obliged to accept any of the bids.

The aggregation combines two scores derived differently. Price is an objective comparative score derived relative to the lowest bid price, expressed as a proportion of the total points available, and then expressed as a proportion of the price weighting. Non-price scores are subjective, derived without comparison to other bids, combined across questions, expressed as a proportion of the total points available, averaged across raters and then expressed as proportion of the non-price weighting.

Despite multiple transformations and differences, when price and non-price scores are aggregated, they are treated as though each point carries equal value. This effectively

creates a nominal exchange rate where compensatory tradeoffs can be made between price and non-price factors: low non-price scores can be compensated by higher price scores and vice versa, low price scores can be offset by higher non-price scores. There is no rational or sound economic basis for the exchange rate disparity established, it is simply a function of the price:non-price weighting.

How much the contracting authority pays is determined by the nominal exchange rate established and small differences in the subjectively derived non-price score can have significant price implications.

Table 3 demonstrates the nominal exchange rate established under different price:non-price weights to help illustrate the implications for the contract price paid due to minor differences in non-price scores. The top two rows of the table show the different combinations of price and non-price weights. Beneath that the rows show the effective difference between price and non-price points under that price:non-price weighting. Points are scaled as one hundred percentage points consistent with percentage weights used.

Table 3: Price/non-price exchange rate

			Weighting									
	Non Price		90%	80%	75%	70%	60%	50%	40%	30%	20%	10%
Scale Points	**Price**		10%	20%	25%	30%	40%	50%	60%	70%	80%	90%
100	**Price Equivalent per Quality Point**		0.9	0.8	0.75	0.7	0.6	0.5	0.4	0.3	0.2	0.1
	Effective Price Premium/ Quality Point	Linear 100	9%	4.0%	3.0%	2.3%	1.5%	1.0%	0.7%	0.4%	0.3%	0.1%
		Prop 100	9.8%	4.1%	3.1%	2.4%	1.5%	1.0%	0.7%	0.4%	0.3%	0.1%

For example, assume a tender with a 10:90 price:non-price weighting receives two bids. Bid A has a non-price score one point higher than Bid B. The price point equivalent is 0.9 or a 9% price difference. To match Bid A's total score, Bid B needs to be 9% cheaper. If Bid A's price is £1 million, Bid B's price will need to be less than £910,000, which is £90,000 cheaper than Bid A under linear pricing or £98,000 cheaper under proportional pricing.

Now, if those two bids are in a competition with a 20:80 price/non-price weighting, Bid B needs to be four per cent cheaper - £40,000 under linear and £41,000 under proportional pricing.

The difference, what we call the price advantage exchange rate is expressed by the formula:

$$PAdv = \big((PLow - PBid) \div PBid\big) \times (PWgt \times 100)$$

Where:

- *PAdv* is the price difference required to match the score of the bid with the higher non-price score
- *PLow* is the price of the lowest priced bid
- *BBid* is the higher priced of the two bids
- *PWgt* is the price weight

These small differences can have significant price implications. Let's assume a tender has a 30/70 price/non-price weighting with two bidders. Bidder A submits a bid of £1 million with a pre-weighted quality score of 55, while Bidder B bids £1.25 million with a pre-weighted quality score of 58 (see Table 4).

Under a linear assessment, Bidder A, as the lowest bidder, receives 100% of the price score, or 30 points. Bidder B's bid is 25% higher, so its price score is reduced to 23 points.

The non-price scores are adjusted to reflect their 70% weighting:

Bidder A's quality score of 55 translates to a weighted non-price score of 39.

Bidder B's quality score of 58 translates to a weighted non-price score of 41.

This results in an eight-point gap in the price score and a three-point difference in the non-price score, leading to a total score difference of 5.4. Ultimately, only the final combined score matters, as the contract is awarded to the bidder with the highest total score.

Table 4 Worked example of bid scoring outcomes under three different price/non-price weights

30% price / 70% nonprice							
	Price	**Price diff as %**	**Score (100-Diff%)**	**Price score (30% weight)**	**Non-price score**	**Non-price score (70% weight)**	**Total score**
Bid A	1,000,000	100	100.0	30.0	55.0	38.5	68.5
Bid B	1,250,000	25	75.0	22.5	58.0	40.6	63.1
Difference	750,000.00		*25*	*7.5*	*-3.0*	*-2.1*	*5.4*
20% price / 80% non-price							
	Price	**Price diff as %**	**Score (100-Diff%)**	**Price score (20% weight)**	**Non-price score**	**Non-price score (80% weight)**	**Total score**
Bid A	1,000,000	100.0	100.0	20.0	55.0	44.0	64.0
Bid B	1,250,000	25.0	75.0	15.0	58.0	46.4	61.4
Difference	750,000.00		*25.0*	*5.0*	*-3.0*	*-2.4*	*2.6*
10% price / 90% non-price							
	Price	**Price diff as %**	**Score (100-Diff%)**	**Price score (10% weight)**	**Non-price score**	**Non-price score (90% weight)**	**Total score**
Bid A	1,000,000	100.0	100.0	10.0	55.0	49.5	59.5
Bid B	1,250,000	25.0	75.0	7.5	58.0	52.2	59.7
Difference	750,000.00		*25.0*	*2.5*	*-3.0*	*-2.7*	*-0.2*

Although Bidder A has a total score six points higher than Bidder B, a six-point change in non-price rating, such as Bidder A getting three less points and Bidder B securing three more points across all non-price questions, could reverse positions and see Bidder B win the contract.

When we take the same scenario under a 20:80 price/non-price weighting, Bidder B only needs to close a small non-price gap of 2.6 points. At a price/non-price rating of 10:90, the scores are tied – one point difference in Bidder B's non-price score would win it the contract. A one-point difference during the non-price evaluation in Bidder B's favour could hand the contract to Bidder B despite the 25% or £250,000 higher price.

The key point from this example is that the tender outcome is highly sensitive to small differences in non-price scores, particularly when the price weight is low. Human rating is imprecise, but the tender process does not include a margin or tolerance for standard error or confidence intervals around total scores. Without a mechanism to question or challenge B's higher score, the contracting authority has two options: reject both bids and start again with a new competition or award the contract to Bidder B and pay 25% (£250,000) more than for Bidder A because of a negligible difference in points.

The magnitude of a potential overspend increases with contract value. If our scenario was for a £1 billion contract, the overspend is £250 million. This is not hyperbole; public contracts of this size do exist.

By using a 10/90 price/nonprice weights to signal that quality is more important than price, contracting authorities amplify the value of non-price factors and risk paying more than necessary for potentially unquantifiable non-price differences.

Steps can be taken to prevent that from happening such as the use of caps and collars (upper and lower limits) on the maximum price a bidder can bid to remain compliant. Bids which exceed these limits could be excluded from the competition. Caps and collars are used in some industries, such as construction. For example, a cap was used in the £1 billion tender for the HS2 Old Oak Common rail project. Bechtel, one of two bidders unsuccessfully challenged the outcome due to the small differences in non-price scores and what it called an abnormally low bid by its competitor (*Bechtel Ltd v High Speed Two (HS2) Ltd*, 2021), but they are not standard practice. Hence, it will not be unusual to receive bids with a large spread of prices. The key question arising from this scenario is how confident are contracting authorities that the evaluator's non-price ratings are calibrated to achieve that degree of both accuracy and precision? Can such a small difference in non-price factors compensate for significant price differences?

Precise or rounded averages

Evaluators rate using whole rating scale integers but aggregation and averaging may result in precise averages which use decimal points. Whether precise or rounded averages are used has a significant effect on the non-price score.

Imagine, for example, two bids have a non-price score difference of 0.2: Bidder A gets 7.6 and Bidder B gets 7.4. If rounding is used the 0.2 gap between bid scores widens to one: 7.6 is rounded up to eight and 7.4 is rounded down to seven. In a scenario with a 20:80 price/non-price weighting - or 20 price points and 80 non-price points - to achieve parity with Bidder A and its one-point lead, Bidder B needs to be four per cent cheaper.

If precise scores are used Bidder B needs to be 1.8% (£18,000) cheaper to win the contract. On a £1million contract one point is potentially worth £90,000 (9% discount). If that discounted price is not available, the contracting authority may pay for the more expensive bid without receiving any additional benefit – that point, or fraction of a point – could be due to rating noise.

The risk of paying too much occurs when non-price ratings are sensitive to small differences in non-price scores between bids and can overcome price disadvantages so that more expensive bids win without discernible benefits.

The effect of moderation and tied scores

An experimental study of the effects of moderation on bid rating (Izatt, 2024c) shows that there is a high probability that individual evaluator ratings will change during moderation. Rating reductions are far more likely than increases; and the size of the reductions is nearly twice as large as the rating decreases. These changes are likely to change the outcome of a competition based on the original averaged ratings (Izatt, 2024d).

However, what matters is the *difference* in non-price scores between bids, not whether those scores are higher or lower. For example, when price and non-price scores are aggregated it will make no difference to the price, or who wins the contract, if the individual non-price scores of 73 (Bid A) and 75 (Bid B) become 54 (Bid A) and 56 (Bid B) after moderation. In both situations there is a two-point non-price difference. Referring again to Table 4, under a 20:80 price/non-price weight Bid A will need to be eight per cent cheaper than Bid B to win (4% for each one-point difference in non-price score).

A change in scores matters if the *difference* between two non-price scores changes. For example, a four-point improvement in Bid A's 73 to a score of 77, and a one point decrease in the Bid B score to 74 widens the difference to three points. Now Bid B needs to be 12% cheaper under a 20:80 price/non-price weight.

Moderation increases the likelihood of ties, or equal scores, for the non-price evaluation (Izatt, 2024c). This is problematic when the highest non-price scoring bids are tied as failure to differentiate on non-price factors means lowest price becomes the differentiator when price and non-price scores are aggregated.

Rounding scores to whole integers also increases the risk of ties on non-price factors, making price the key determinant.

The risk of paying too little

The risk of paying too little occurs when the lowest price against which all prices are assessed is accepted at face value without verification or reference to a target or should cost model and non-price scores are tied. When this happens, the bids are not differentiated on quality and price becomes the key determinant of the competition.

The lowest price offered has a significant impact in determining the outcome of a contract award. Acceptance of the lowest price without verification, ideally against a target price or should cost price validated as realistic and achievable, leaves the buyer open to paying too little. By too little we mean the contract cannot be delivered or will be delivered to an inferior quality.

How low is too low? Abnormally low tenders are defined as those below market value – meaning what the market would expect to pay. Faced with a spread of prices from the seven bidders in our scenario ranging from £100,000 to £300,000, are the lowest bids abnormally low, or are the two tops bids abnormally high?

The new UK Procurement Act (2023, s 19(4)) allows a contracting authority to disregard a tender from a supplier that offers a price that the contracting authority considers to be abnormally low for the performance of the contract. Before disregarding a suspected abnormally low price the contracting authority must notify the supplier that it considers the price to be abnormally low and give the supplier reasonable opportunity to demonstrate "to the contracting authority's satisfaction" that it will be able to perform the contract for the price offered. No definition or guidelines for what would trigger an abnormally low notice is provided.

This mirrors the regulatory regime under the EU Public Contracts Regulation (2015) where it is mandatory for a contracting authority to seek an explanation or clarification of the price or costs if they "appear" to be too low[3] before rejecting it. The contracting authority can only reject it if it has requested in writing an explanation of the offer or of those parts which it considers contribute to the offer being abnormally low; taken account of the evidence provided in response to a request in writing; and subsequently verified the offer or parts of the offer being abnormally low with the economic operator. Under the 2006 regulations this investigation was non-mandatory and there were few contested cases in the UK.

One example of efforts to define an abnormally low tender is the Northern Ireland Government 's policy for dealing with suspected abnormally low tenders in the construction industry. It calculates what it calls an adjusted average – the average of all price bids excluding the highest price bid. The price should be no greater than 15% lower than the adjusted average plus a proximity margin of 1% with a minimum of £1,000 and maximum of £100,000.

In our scenario (see Table 5) where six bid prices are submitted, the adjusted average for the bids is £147,500. The adjusted average after removing the most expensive bid is £147,500. The lowest acceptable bid is £125,362 so Bids A and B would be deemed abnormally low, and the contracting authority would need to challenge their bids. If they were removed Bid F, which received no points under linear pricing before, now gets price points. The highest price bid still fails to get price points even with Bids A and B removed.

[3] Regulation 69(1) of the Public Contracts Regulations 2015 ("**PCR 2015**").

Table 5: Sample Abnormally Low Tender (ALT) calculation

Bidder	Bid Price
A	£100,000
B	£120,000
C	£140,000
D	£150,000
E	£175,000
F	£200,000
G	£300,000

Average	£169,286
Adjusted average	£147,500
Lowest acceptable	£125,375.00
Range +1%	£126,628.75
Range -1%	£125,362.46

Discussion

A substantial body of extant literature provides evidence of noise and bias in judgement (Kahneman, Sibony and Sunstein, 2021). While variability in judgement has been studied in areas such as exam marking, university admissions and court sentencing (Kahneman, Sibony and Sunstein, 2021), it has not been investigated in public procurement where judgement plays a significant role in the deployment of billions of pounds of taxpayer money.

The purpose of this paper was to address this gap and demonstrate the effects and implications of noise, or variation, in non-price ratings, on contract award outcomes under different price and non-price weights and contract values.

By considering the combination of subjective non-price and objective price scores under different weights, this analysis identifies the potential for paying too much for no added value because of the sensitivity of non-price evaluation to unexplained or random variations in evaluator ratings.

The risk of overpaying increases as the weighting of non-price factors increases relative to price. It could be argued that this reflects the buyer's intention to focus on quality rather than price, but it does not satisfy the objective of public procurement which is to deliver value for money. Setting a higher non-price weight does not in itself guarantee the bids with the best quality wins as differences in scores may be the result of random variability due to inconsistency in how an evaluator rates, and inconsistency across raters.

Weighting is not the only problem, there are two others. The first is the number of transforms used and the second is the compensatory aggregation. Tenders are awarded based on judgements which are transformed through averaging, weighting, rounding and aggregation with other scores which have also gone through multiple transformations. While the approach superficially looks robust, the outcomes are based on spurious precision (Bobko, Roth and Buster, 2007). This paper has demonstrated the effect these can have on final bid scores by cancelling out, increasing or decreasing differences between bid scores.

The compensatory trade-off that occurs when the objectively derived (using comparison) price and the subjectively derived (without comparison) non-price scores are combined is problematic and inappropriate. It treats price and nonprice scores as a common currency when the two are combined (Greco et al., 2019). This would be like having a wallet with a mix of currency denominations but only considering the face value rather than an exchange rate so that one euro, one peso and one British pound would have the same value.

Calculating the exchange rate that occurs between price and non-price points demonstrates how the scale of the problem increases as buyers signal that quality is more important than price by using a 90:10 weight. Conversely, while a 50:50 price/non-price weighting might be perceived as favouring lowest price, this weighting that removes the imbalance in a compensatory model.

This matters because small differences can change the outcome of tenders. Bid evaluators will be unaware of contract price implications when they award a bid one point more than another or agree in moderation to change their rating either up or down. That’s because price and non-price evaluation are assessed independently of each other. Would evaluators rate differently if they knew that a one-point difference in score is worth £23,000 or £90,000? Could they identify the reason for a one-point difference between bid quality scores?

Noise has been found to increase with the complexity of a specification (Izatt, 2024b). Rater consistency is less likely to be an issue when there is a correct answer or benchmark against which the object to be evaluated can be compared. For example, when choosing between bids which all comply with a technical specification, with all things being equal except price, it makes sense to accept the lowest price. As public authorities have moved towards increased use of performance or output specifications, all things being equal is no longer the case. Different bidder approaches, methods, designs and materials result in a lack of commensurability and measurability (Walasek and Brown, 2023) and evaluation is increasingly subjective. However, the underlying evaluation process has not changed in response to the change from simple to complex contracting.

Implications

The purpose of this paper was to demonstrate the consequences of contract award outcomes for achieving value for money due to the current compensatory bid scoring approach. A rethink is required to hit the Goldilocks sweet spot between paying too much and too little to get value for money (see Table of recommendations).

Supporters of the current process may argue that it prevents bribery and corruption by reducing the influence of decision makers on the outcome. However, this overlooks the presence of human bias in judgement. The usual fairness defense that the process is fair because all bids are treated the same is inadequate. Just because everyone is treated the same does not make it fair if the underlying process is unfair (Rawls, 1971).

Given the scale of public sector outsourcing a review of bid evaluation is overdue. Changes are required to both the design and evaluation of price and non-price evaluation.

When criticisms are made of an existing process, the predictable question is what do you suggest we do about it? Should changes be made by policy makers or practitioners? We have some suggestions.

Recommendations

Starting with policy makers, the current tender evaluation policy appears to offer few restrictions to change. All bidders should be treated fairly. Bidders should be told before bidding how their bids will be evaluated, against what criteria, and how contracts will be awarded. Bidders have the right to feedback about their bid and the right to challenge the outcome of a tender. None of this needs to change.

There are a few recommendations for policy and guidance change:

- Redefine fairness, which was narrowly defined in the EU procurement regulations, to include the requirement for the evaluation process to be fair.
- Rethink the government guidance commitment to separate evaluation. Joint evaluation rating happens whether you expect it to or not. Stating that separate evaluation should be used in moderation may leave contracting authorities open to breaches of stated procedure when raters inevitably compare bids. Why run the risk?

Changes, overall, would appear to be in the hands of practitioners. Based on our research findings in this and our other papers, recommendations for practitioner changes to how bids are evaluated include:

- Increase the rating scale length to 10 points and standardise its use across tenders. It should reduce ties and increase bid differentiation. The commonly used five-point rating scale is too short.
- Train evaluators so they understand the consequences of noise in award outcomes.
- Use more evaluators. Averaging ratings assumes a large pool of ratings. Averaging two or three ratings hides disagreements. Using more raters makes the average more meaningful and investigate differences in ratings.
- Avoid rounding scores to the nearest whole integer to avoid undue advantage or disadvantage to bidders. Rounding could result in paying more than necessary.
- Assess quality first. Remove poor quality bids by introducing cut scores for quality before assessing their price. The cut score determines the minimum acceptable quality. The winning bid should meet or exceed the cut score on all criteria. This removes the risk of low-quality cheaper bids winning, but retains the possibility of a high-quality cheaper bid winning.
- Replace the compensatory composite score with a non-compensatory model in which price and non-price scores are not combined. Bids should pass a cut score on all categories (non-price and price) to stay in the competition. Cut scores should be sufficiently high to ensure both the desired quality and the desired price are achieved.
- Introduce an additional holistic assessment to the current analytical assessment. Once evaluators have rated all sections, ask them to provide an overall rating.
- Ask evaluators to rate their confidence in their rating, and their confidence in their ability of the rater to deliver the contract. The same approach could also be used in moderation.
- Introduce confidence levels or a buffer around total scores – e.g. at least 4 points difference between bids - so that contracts are not awarded based on small differences in total scores which could arise from noise rather than substantive differences.
- If the two highest scoring bids differ by less than the required amount, trigger an additional review of the bids. For example, that additional review could be a holistic assessment of ratings for individual bid sections. Or choose.
- Use holistic evaluation (overall assessment of the bid) to test the outcome from the mechanical evaluation (aggregating scores from different sections).
- Shift from process to outcome accountability, appointing individuals who are responsible for the outcome. Supplementing judgement with choice may raise concerns about the

potential for bribery and corruption. However, it would improve transparency and contracting authority accountability.

Limitations

The conclusions from this study apply to tenders without negotiation. Where negotiation is used then the differences identified may be clarified and changes may be negotiated. However, in both cases, the contracting authority is still required to set out the process by which bids will be evaluated and under current regulations those weights and criteria cannot be altered once published.

We have focused on the possibility of overpaying due to negligible differences in non-price scores although of course paying too little is also a risk. Both are problematic. Paying more than necessary is a waste of taxpayer money particularly at a time when the public is constantly reminded of the existence of an ever-growing fiscal black hole. Paying too little is a problem if it means contracts are awarded to poor quality bids because their low price allows them to overcome poor quality scores. As with overpaying, paying too little can also waste taxpayer funds and incur more costs if a service has to be replaced or bailed out, such as in the case of Carillion.

This paper merely illustrates the potential effects of the current tender evaluation practice and does not constitute a legal position or description of any actual contract award.

Conclusion

The UK Government awards £400 billion of contracts annually, much of it through a procedure which creates the potential to overpay unnecessarily for no discernible benefit due to the sensitivity of tender outcomes to small differences in non-price ratings. By failing to adequately account for noise and bias in subjective ratings in the bid scoring procedure, value for money is compromised. To achieve value for money in public procurement, it is time to retire a mechanism based on lowest price for one capable of finding the best bid on both quality and price.

References

Arrowsmith, S. (2005) *The Law of Public and Utilities Procurement.* 2nd edition. Sweet & Maxwell.

Barkaoui, K. (2011) ‘Effects of marking method and rater experience on ESL essay scores and rater performance’, *Assessment in Education: Principles, Policy and Practice*, 18(3), pp. 279–293. Available at: https://doi.org/10.1080/0969594X.2010.526585.

Bechtel Ltd v High Speed Two (HS2) Ltd (2021). LNUK. Available at: https://plus.lexis.com/api/document?collection=cases-uk&id=urn:contentItem:624K-DGK3-GXFD-82WP-00000-00&context=1001073.

Bobko, P., Roth, P.L. and Buster, M.A. (2007) ‘The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis’, *Organizational Research Methods*, 10(4), pp. 689–709. Available at: https://doi.org/10.1177/1094428106294734.

De Boer, L., Labro, E., & Morlacchi, P. (2001). A review of methods supporting supplier selection. *European Journal of Purchasing and Supply Management*, *7*(2), 75–89. https://doi.org/10.1016/S0969-7012(00)00028-9

Bruno, G., Esposito, E., Genovese, A., & Passaro, R. (2012). AHP-based approaches for supplier evaluation: Problems and perspectives. *Journal of Purchasing and Supply Management*, *18*(3), 159–172. https://doi.org/https://doi.org/10.1016/j.pursup.2012.05.001

Cheaitou, A., Larbi, R. and Al Housani, B. (2019) ‘Decision making framework for tender evaluation and contractor selection in public organizations with risk considerations’, *Socio-Economic Planning Sciences*, 68. Available at: https://doi.org/10.1016/j.seps.2018.02.007.

Contracts (Applicable Law) Act (1990).

Dawes, R.M. (1979) ‘The robust beauty of improper linear models in decision making’, *American Psychologist*, 34(7), pp. 571–582. Available at: https://doi.org/10.1037/0003-066X.34.7.571.

Dotoli, M., Epicoco, N. and Falagario, M. (2020) ‘Multi-Criteria Decision Making techniques for the management of public procurement tenders: A case study’, *Applied Soft Computing*, 88, p. 106064. Available at: https://doi.org/10.1016/J.ASOC.2020.106064.

Falagario, M. *et al.* (2012) ‘Using a DEA-cross efficiency approach in public procurement tenders’, *European Journal of Operational Research*, 218(2), pp. 523–529. Available at: https://doi.org/10.1016/j.ejor.2011.10.031.

Gigerenzer, G. and Goldstein, D.G. (1996) ‘Reasoning the fast and frugal way: Models of bounded rationality’, *Psychological Review*, 103(4), pp. 650–669. Available at: https://doi.org/10.1037/0033-295X.103.4.650.

Goldstein, W.M. and Einhorn, H.J. (1987) *Expression Theory and the Preference Reversal Phenomena*, *Psychological Review*.

Government Commercial Function (2021) *Bid Evaluation*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/987130/Bid_evaluation_guidance_note_May_2021.pdf.

Greco, S. *et al.* (2019) 'On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness', *Social Indicators Research*, 141(1), pp. 61–94. Available at: https://doi.org/10.1007/s11205-017-1832-9.

Grupp, H. and Mogee, M.E. (2004) 'Indicators for national science and technology policy: How robust are composite indicators?', *Research Policy*, 33(9), pp. 1373–1384. Available at: https://doi.org/10.1016/j.respol.2004.09.007.

Grupp, H. and Schubert, T. (2010) 'Review and new evidence on composite innovation indicators for evaluating national performance', *Research Policy*, 39(1), pp. 67–78. Available at: https://doi.org/10.1016/j.respol.2009.10.002.

Highhouse, S. (2008) 'Stubborn Reliance on Intuition and Subjectivity in Employee Selection', *Industrial and Organizational Psychology*, 1(3), pp. 333–342. Available at: https://doi.org/10.1111/j.1754-9434.2008.00058.x.

Ho, W., Xu, X., & Dey, P. K. (2010). Multi-criteria decision making approaches for supplier evaluation and selection: A literature review. *European Journal of Operational Research*, *202*(1), 16–24. https://doi.org/10.1016/j.ejor.2009.05.009

Hodges, B. (2013) 'Assessment in the post-psychometric era: Learning to love the subjective and collective', *Medical Teacher*, 35(7), pp. 564–568. Available at: https://doi.org/10.3109/0142159X.2013.789134.

House of Commons Public Administration and Constitutional Affairs Committee (2019) *After Carillion: Public Sector Outsourcing and Contracting*. Available at: https://publications.parliament.uk/pa/cm201719/cmselect/cmpubadm/748/748.pdf.

Hsee, C.K. (2019) *Attribute evaluability: Its implications for joint-separate evaluation reversals and beyond*, *Choices, Values, and Frames*. Available at: https://doi.org/10.1017/CBO9780511803475.032.

Izatt, J. (2024a). Accountability: Is it overrated? In Unpublished manuscript. Part of doctoral thesis, University of Warwick.

Izatt, J. (2024b). Evaluation mode and evaluability: Which drives tender rating variability? In Unpublished manuscript. Part of doctoral thesis, University of Warwick.

Izatt, J. (2024c). On second thought, it's worse: the effect of moderation on bid ratings and tender outcomes. In Unpublished manuscript. Part of doctoral thesis, University of Warwick.

Kahneman, D., Sibony, O. and Sunstein, C.R. (2021) *Noise A Flaw in Human Judgement*. London: William Collins.

National Audit Office (2018) *Investigation into the government's handling of the collapse of Carillion*. Available at: https://www.nao.org.uk/reports/investigation-into-the-governments-handling-of-the-collapse-of-carillion/.

National Audit Office (2024) *Efficiency in government procurement of common goods and services*. Available at: https://www.nao.org.uk/reports/efficiency-in-government-procurement-of-common-goods-and-services/.

Niewerth, S., Vogt, P., & Thewes, M. (2022). Tender evaluation through efficiency analysis for public construction contracts. *Frontiers of Engineering Management*, *9*(1), 148–158. https://doi.org/10.1007/s42524-020-0119-z

Parducci, A. (1965) 'Category judgment: A range-frequency model', *Psychological Review*, 72(6), pp. 407–418. Available at: https://doi.org/10.1037/h0022602.

Parducci, A. and Wedell, D.H. (1986) 'The Category Effect With Rating Scales. Number of Categories, Number of Stimuli, and Method of Presentation', *Journal of Experimental Psychology: Human Perception and Performance*, 12(4), pp. 496–516. Available at: https://doi.org/10.1037/0096-1523.12.4.496.

Procurement Act (2023).

Stevens, S.S. (1946) 'On the theory of scales of measurement', *Science*, 103(2684), pp. 677–680. Available at: https://doi.org/10.1126/science.103.2684.677.

Stewart, N., Chater, N. and Brown, G.D.A. (2006) 'Decision by sampling', *Cognitive Psychology*, 53(1), pp. 1–26. Available at: https://doi.org/10.1016/j.cogpsych.2005.10.003.

Sturgess, G. (2017) *Just another paperclip? Rethinking the market for complex public services*. Available at: https://golab.bsg.ox.ac.uk/documents/Just_Another_Paperclip_-_Rethinking_the_Market_for_Complex_Public_Services.pdf (Accessed: 10 June 2022).

UK Government (no date) *Contracts Finder*. Available at: https://www.gov.uk/contracts-finder (Accessed: 22 November 2024).

Vlaev, I. *et al.* (2011) 'Does the brain calculate value?', *Trends in Cognitive Sciences*, 15(11), pp. 546–554. Available at: https://doi.org/10.1016/j.tics.2011.09.008.

Walasek, L. and Brown, G. (2023) 'Incomparability and Incommensurability in Choice: No Common Currency of Value?', *Perspectives on psychological science : a journal of the*

Association for Psychological Science, p. 17456916231192828. Available at: https://doi.org/10.1177/17456916231192828.

Watt, D. J., Kayis, B., & Willey, K. (2010). The relative importance of tender evaluation and contractor selection criteria. *International Journal of Project Management*, *28*(1), 51–60. https://doi.org/10.1016/j.ijproman.2009.04.003

Weber, C. A., Current, J. R., & Benton, W. C. (1991). Vendor selection criteria and methods. *European Journal of Operational Research*, *50*(1), 2–18. https://doi.org/10.1016/0377-2217(91)90033-R

World Trade Organization (2024) *General Procurement Agreement*. Available at: https://www.wto.org/english/tratop_e/gproc_e/gp_gpa_e.htm (Accessed: 5 September 2024).