Quantity Choice in Unit Price Contract Procurements

Svante Mandell and Fredrik Brunes

Working Paper 2012:02

Section for Building and Real Estate Economics
Department of Real Estate and Construction Management
School of Architecture and the Built Environment
Royal Institute of Technology
Abstract

Unit Price Contracts is a procurement approach commonly used for construction projects. We develop an analytical model to study the optimal procurement quantity and monitoring intensity when the required quantities are uncertain. The optimum involves trading-off risking paying for more units than necessary and conducting costly renegotiations and/or monitoring. The paper adds to the understanding of both optimal behavior in procurements and the presence of cost-overruns. In particular, deliberately procuring low quantities, and thereby facing a high risk of cost-overruns, is sometimes optimal as it minimizes the expected total cost.

Keywords: Unit price contracts; procurement; construction; cost-overruns

JEL codes: H54; H57; D44
1.0 Introduction

Construction projects, e.g., infrastructure projects, may be procured in a number of different ways. The procurement or contract forms can be divided in different dimension. One fundamental division is whether the client or the contractor makes the detailed design (Design-Bid-Build contracts vs Design Build contracts). Another important question is whether construction and operation/maintenance should be procured as in PPP-projects or procured separately which has been the dominating form in most countries. Contracts can also differ in the ways that the contractor is paid.

In this paper the focus is on traditional Design-Bid-Build contracts which still seem to dominate in most countries. A Design-Bid Build contract can be a fixed price contract but in order to reduce the risk for the contractor the contract can contain what in Sweden is called “fixed” and “variable quantities”. In the first case the contractor is paid a fixed amount independently of how many units actually are needed. In the second case the contractor announces a fixed price per unit in their offer, and then the contractor is paid according to the “actual” number of units needed. In order to compare bids the client initial announces a predicted quantity that is used to calculate the expected price of each bid.

In many countries, the prevailing approach in practice seems to be Design-Bid- Build contracts including components where Unit Price Contracts (UPC) is used. This paper focus only on the “variable quantity” part of the contract.In such a UPC, the procurer, e.g. a national road administrator, specifies the amounts of each activity, e.g., the amount of gravel to be

---

1 Love (2002) notes that “traditional lump sum” procurements (to which UPC belongs) dominate in many commonwealth countries. In Sweden, ~ 90% of the road investments between 2000 and 2009 were procured under UPC, Mandell and Nilsson (2010).
removed, and lets the agents bid on unit prices. Typically, the agent with the lowest total bid – summing over all amounts times the bidding prices – wins the procurement.

This theoretical article addresses the optimal behavior of the procurer, henceforth the principal, in UPC procurements. In particular, it addresses what amount of an activity to specify in the contract when the actual amount required is uncertain. Previous work dealing with UPC more or less implicitly assumes that the principal in the UPC-contract will enter the quantity of work that the client think will be needed. For instance, Ewerhart and Fiesler (2003) states that in a UPC “the buyer estimates the quantities of the respective input factors that will be needed to accomplish the task. Then the buyer publicly announces her estimates […]”. We will, by the means of a simple model, show that this notion is not correct. Rather, there are cases in which the principal should – in order to minimize her expected total costs – in the procurement document announce a quantity exceeding the estimated or expected one and other cases in which the announced quantity should be lower. More importantly, the model will provide us with an intuitive understanding for the mechanisms at work.

This article is akin to a literature on optimal behavior among bidders in UPC procurement that addresses strategic bidding behavior when the bidding agents have superior information. The agent may exploit their information advantage by skewing their bids, often referred to as unbalanced bidding. This means that the bidder will announce a low price for amount that they think is overestimated in the procurement document and announce a high price for activities that the bidder think is underestimated in the procurement document. The underlying information asymmetry may be that the agent is better informed about the actual, ex post, amounts of individual tasks. This case is investigated by Atey and Levin (2001) and Bajari et.al. (2007). A similar situation may occur when the agent is better informed about her
own type, e.g., skill, as studied in the aforementioned Ewerhart and Fieseler (2003). These articles model the bidding agents’ behavior in UPC auctions whereas the present article models the procurer’s behavior. Thus, the present article is a step towards a unified model which allows for strategic behavior of both procurer (client) and bidders (contractors).

As noted, a central outcome of our model is that the quantity to announce may deviate, upwards or downwards, from the expected quantity actually required. Consequently, our model adds to the literature on cost overruns. That cost overruns are frequently occurring in construction projects seems to be an established fact in the literature, Flyvbjerg et.al. (2003), Odeck (2004), and Berechman and Chen (2011). According to Priemus et al (2008), cost overruns for large infrastructure projects of between 50 and 100 per cent are common. Furthermore, the forecasts of costs have not improved over the last 70 years.

Priemus et.al. also provide a couple of plausible explanations for systematic miscalculation of costs leading to cost underestimation; bad forecasting due to technical problems and/or calculation problems, that the project change shape during the construction phase, and that planners intentionally perform a forecast to support an already decided project. Our model suggests one additional reason. Namely that the procurer, in some situations, contracts on a quantity that she knows is most probably less than required and, thus, total costs *ex post* will most probably exceed the contracted sum. The interesting – and perhaps somewhat paradoxical – result is that this can be optimal as it in some situations in order to keep the *expected* total costs at a minimum. It is important to underline that this possible rational mechanism behind cost-overruns is additional to the kind of factors discussed by e.g. Flyvbjerg where political/strategic factors are central.
Ganuza (2007) and Gaspar and Leite (1989/90) are related to the present article as both develop models on procurement in which cost-overruns are likely to occur in optimum. These studies focus on different aspects of procurement than our model. The former shows that the procurer, in optimum, should underinvest in design specification. The reason is that an exact design will decrease competition among bidders, which results in a large share of the rents being captured by the winning bidder. The latter provides a model in which each bidder has an imperfect signal about the cost of finalizing the project. As the lowest bid will win, a selection bias problem emerges. This results in a high risk for cost-overruns.

It should be noted that there are different definitions of the concept cost-overrun. The definition here, as in Ganuza (2007) and, to some extent, Gaspar and Leite (1989/90), would be actual total costs minus contracted sum. Priemus et al. (2008, page 125) states the definition as “Actual cost minus forecasted cost” where forecasted cost is defined as “the estimate made at the time of decision to build, or as close to this as possible if no estimate was available for the decision to build”. The contract sum is probably not a good approximation of the latter. The remaining article is structured as follows: The next section introduces the model and leads up to two first-order conditions. The characteristics of these are analysed in section 3. This section is divided into two subsections; the first in which the principal is unable to monitor the agent, and the second in which monitoring is possible. Section 5 concludes.

2.0 The model

We will not model the bidding procedures. Rather, we assume that the winning bid covers the agent’s marginal cost associated with each activity with some margin. Consequently, the agent always gains from conducting one extra unit of the activity and has no incentive to carry
out less of an activity than what is specified in the contract. A crucial question is how easy it is for the principal/client to observe the actual quantities. An extreme case, is if the principal cannot observe (very high cost of monitoring) how many units are actually required, the amounts specified in the contract will then serve as a lower threshold for the amount that the agent will carry out\(^2\). There are two important aspects to note: First, we do not allow the agent to lie, \textit{i.e.}, report a larger number of units than actually conducted, in our model. The agent will carry out \(q\) units, even if the project could have been finalized with less. This maximizes the agent’s profit given the assumption that the unit price exceeds marginal costs. If this assumption is relaxed, such that at some point the marginal cost becomes larger than the unit price, the agent may not conduct more work if not necessary. This could be incorporated in the model, but it would not yield any additional insights. Second, we assume that the contracted amount is a strict threshold. In real life this may not be the case. Rather, the contract may allow for more units than procured to be conducted. However, at some point renegotiation will be necessary\(^3\). This only implies that the renegotiation point is moved from the contracted amount to some larger amount, \textit{i.e.}, the contracted amount plus 25 per cent. The subsequent model and its implications are still valid.

For the sake of this presentation it suffices to focus on one activity. Let us denote the lowest amount of this activity required to complete the project by \(Q\). It is easy to expand the model to

\(^2\) For this to be true there must be nothing else to gain from conducting less of an activity than the contract states. In particular, there may be no reputational effects. That is, the agent’s behaviour in this contractual relationship must not influence the probability of winning future contracts.

\(^3\) Or else, the costs of any contract may be prohibitively high. In Sweden, the typical contract (as defined by a standard denoted AB04, chapter 6 §6) states that if the actual number of units differs, upwards or downwards, from the contracted number of units by 25 per cent or more the contract (the unit prices) should be renegotiated if either party requests it.
include several activities, but it will not add to the understanding of the problem. In the procurement stage, i.e., *ex ante*, \( Q \) is not fully known\(^4\). However, it seems reasonable that the principal has a prior but uncertain estimate of \( Q \). Let us assume that \( Q \) is uniformly distributed on the interval \( Q_{\text{low}} \) to \( Q_{\text{high}} \). A uniform distribution is not very realistic but used here to facilitate the presentation. In a recent paper, Berechman and Chen (2011) find that the data on cost overruns they study is best described using a Cauchy distribution. Potential impacts on our results of allowing for a more bell shaped, rather than a uniform, distribution will be briefly addressed in a later section.

Given this information, the principal specifies an amount in the UPC. Let us denote this amount \( q \). The bidding process yields a winning per unit price for \( q \), which we denote \( p \). Clearly, when deciding what amount to procure the resulting price is not known. *Intuitively, and as will be shown subsequently, the principal’s belief about the emerging per unit price will influence his choice of \( q \).* To capture this, we assume that the principal knows that the emerging price will be in the uniform\(^5\) interval \( (p_{\text{low}}, p_{\text{high}}) \).

Due to the uncertainty regarding required quantity, it may be the case that \( q \) is not sufficient to produce the project, *i.e.*, it may be the case that \( q < Q \). Whether or not this situation occurs is not known at the procurement stage, but becomes evident to the agent during the construction phase.

If it turns out to be the case that \( q < Q \) the principal and the agent may renegotiate the contract in order to finalize the project. This may be associated with a renegotiation cost, denoted \( R \).

\(^4\) This may, for instance, be due the exact characteristics of the rock may be unobservable prior to the project has started.

\(^5\) The crucial assumption is that the distribution is symmetric around its expected value.
We restrict our attention to $R \geq 0$. There is also an upper limit such that the cost of 
renegotiating the contract will not exceed the entire possible gain from renegotiating\(^6\). The 
result of the renegotiation is a price per unit for the remaining amount of the activity required 
to finalize the project, $p_R \geq p$.\(^7\) Let $p_R$ be equal to $p + \gamma$, where $\gamma$ is referred to as a mark-up. 
For simplicity, we assume that at the renegotiation stage the true $Q$ is observable for both 
parties. That is, there will only be one renegotiation as it is then known (with certainty) that 
the remaining amount is $Q - q$.

When $q \geq Q$, the principal, as mentioned above, may not be able to directly observe $Q$. 
However, it seems realistic to assume that the principal may invest in monitoring, which 
provides a signal about the actual lowest amount needed and carried out, $Q$, even without 
renegotiation taking place. We model this by introducing a monitoring variable, denoted $\alpha$, set 
by the principal at the procurement stage. It may take on values from zero to one and captures 
the probability that the principal observes, and then pays for, the correct quantity\(^8\). The 
expected quantity the principal pays for in the absence of renegotiation then becomes 
$\alpha Q + (1 - \alpha)q$. Monitoring is costly. Denote the monitoring cost $C(\alpha)$, where both $C'(\alpha)$ and 
$C''(\alpha)$ are strictly positive.

---

\(^6\) Then, it is obviously better to procure the maximal amount of the activity, i.e., $q=Q_{\text{high}}$, and thereby setting the probability of having to renegotiate to zero. When the principal cannot monitor the agent, this limit is given by, $R \leq (Q_{\text{high}} - Q_{\text{low}}) p$.

\(^7\) That is, we disregard the case where the renegotiated price is less than the initial one. The rational for this is 
that it seems unlikely that the principal would receive a better deal once locked into an agreement with a given 
agent.

\(^8\) As noted above, the agent will not lie about the quantity used. However, she may conduct more work than 
actually needed. Monitoring provides the principal with a signal about the lowest quantity needed to finalize the
 project.
Figure 1 summarizes the timing of the model. In the first stage, the principal decides on how many units to announce in the procurement document, \( q \), and the level of monitoring, \( \alpha \). This is followed by the procurement, which will establish the winning per unit price. The next step is the construction phase under which the true \( Q \) is realized. This is directly observable by the agent. The principal receives a signal about \( Q \), whose accuracy depends on \( \alpha \). If \( q \) is sufficient, i.e., \( q \geq Q \), payments are made according to the contract and the game ends. If not, renegotiations are needed.

![Diagram]

**Figure 1.** The sequence of the game.

Using the assumptions above, the principal’s total cost, \( TC \), is given by

\[
TC = \begin{cases} 
(p(\alpha Q + (1 - \alpha)q) + C(\alpha)) & \text{when } q \geq Q \\
(pq + R + (p + \gamma)(Q - q) + C(\alpha)) & \text{when } q < Q
\end{cases}
\]

A key factor is that when there is no monitoring, i.e., \( \alpha = 0 \), \( TC \) will always equal \( p \cdot q \) for realizations below \( q \). The principal thus pays for \( q \) units even though the project could have been carried through using only \( Q < q \) units, i.e., \( TC \) are constant for realizations between \( Q_{low} \) and \( q \). If the principal invests in monitoring, i.e., \( 0 < \alpha \leq 1 \), the expected \( TC \) for realizations below \( q \) increases in \( Q \). Another key factor is the cost of renegotiations, which show up in two different ways; first through a shift in the \( TC \)-function at \( q \) (if there is a renegotiation cost, i.e., if \( R > 0 \)), and, second, through the slope of the \( TC \)-function above \( q \) (if the renegotiation results in a higher per unit price, i.e., if \( \gamma > 0 \)).
Neither party knows the actual $TC$ prior to the construction phase. Rather, the principal strives to minimize the expected total cost, which may be written as

$$E\{TC\} = \frac{1}{p_{high} - p_{low}} \int_{p_{low}}^{p_{high}} \left( \frac{1}{Q_{high} - Q_{low}} \left( \int_{q_{low}}^{q} p(aQ + (1 - a)q) \, dq + \int_{q}^{Q_{high}} (pq + R + (p + \gamma)(Q - q)) \, dq \right) \right) \, dp + C(\alpha) \quad (1)$$

The first integral is due to the uncertainty in emerging price prior to conducting the procurement. The first integral inside the brackets captures realizations under which $q$ is sufficient. The second integral captures those where additional activity is required.

The principal’s optimization problem amounts to choosing $q$ and $\alpha$ to minimize $E\{TC\}$. Taking account for that the symmetric distribution of $p$ implies that $p_{low} + p_{high} = 2\bar{p}$, where $\bar{p}$ denotes the expected price, we may use (1) to derive a first order condition for the optimal quantity to procure as $^9$

$$q^* = \frac{pQ_{low}(1-\alpha)+R+Q_{high}\gamma}{p(1-\alpha)+\gamma} \quad (2)$$

Also from (1) we reach the following condition that constitutes the optimal level of monitoring.

$$C'[\alpha] = \frac{p(q-Q_{low})^2}{2(Q_{high}-Q_{low})} \quad (3)$$

The next section contains the analysis of the characteristics of these conditions for optimality.

$^9$ The uncertainty surrounding the price does not enter into the expressions as long as it, as is assumed here, is symmetrically distributed around the expected value.
3.0 Analysis

From (2) and (3) it is evident that both the optimal quantity to procure and the optimal monitoring level are influenced by several variables. The aim of this section is to analyze and provide an intuitive understanding for how these variables affect the outcome in optimum.

The principal decides on both the quantity to present in the procurement documents and the monitoring level before the procurement process starts. The optimal quantity, see (2), depends on the monitoring level, and the optimal monitoring level, see (3), depends on the chosen quantity. That is, the two optimization problems are to be solved simultaneously. It is possible to do this analytically, but the resulting first order conditions become cumbersome. In order to facilitate the intuitive understanding, we therefore start by focusing on \( q^* \) in isolation by assuming that no monitoring is conducted, i.e., we set \( \alpha \) to zero\(^{10}\). We will return to optimal monitoring subsequently.

3.1 The case with no monitoring (\( \alpha = 0 \)).

If the principal has no possibility to invest in monitoring, the first order condition relevant for the quantity, (2), simplifies to

\[
q_0^* = \frac{\bar{p}Q_{low} + \gamma Q_{high} + \bar{R}}{\bar{p} + \gamma}
\]

(2')

Where the subscript denotes that \( \alpha = 0 \). We start with the influence from the renegotiation cost. Differentiating (2’) with respect to \( R \) yields:

\[
\frac{\partial q_0^*}{\partial R} = \frac{1}{\bar{p} + \gamma} > 0
\]

(4)

\(^{10}\) This would be motivated in optimum if monitoring were prohibitively costly.
The denominator of (4) is the expected price after the renegotiation. Even if we would allow for a negative price mark-up, we do not allow for this price to be negative. Thus, (4) is positive and the optimal quantity increases (linearly) in the renegotiation cost, see Figure 2a. This makes intuitive sense. By increasing the procured quantity, the risk of ending up in a situation where this quantity is insufficient, i.e., \( q < Q \), becomes smaller. On the other hand, by increasing \( q \), the cost incurred when \( q \) turns out to be sufficient, i.e., \( q \geq Q \), becomes larger. This illustrates the fundamental trade-off faced by the principal; a larger \( q \) reduces the risk for costly renegotiations, but also increases the costs for outcomes where renegotiations are not needed. What (4) shows is that if the renegotiation costs are higher, then the principal is willing to increase the costs incurred when \( q \geq Q \) to reduce the risk of renegotiations.

Applying the same logic on the price mark-up, \( \gamma \), would suggest that the optimal procured quantity should increase if the mark-up is increased. To see this, we differentiate (2') with respect to \( \gamma \), which yields

\[
\frac{\partial q_0}{\partial \gamma} = \frac{\bar{p}(Q_{\text{high}}-Q_{\text{low}})-R}{(\bar{p}+\gamma)^2} > 0
\]

Equation (7) is indeed positive\(^{11}\). Thus, if the price mark-up increases, the principal increases the contracted quantity and thereby reduces the risk of having to pay the higher renegotiated price, see Figure 2b.

If the expected per unit price increases, the principal will reduce the procured quantity, as seen from differentiating (2’) with respect to \( \bar{p} \) which yields

---

\(^{11}\)This is true as long as \( R \leq (Q_{\text{high}}-Q_{\text{low}}) \bar{p} \) which must be the case as otherwise the contractor will put \( q = Q_{\text{High}} \) and thereby ruling out any risk for renegotiation

13
\[
\frac{\partial q_0}{\partial p} = -\frac{R^+ (q_{\text{high}} - q_{\text{low}}) y}{(p + y)^2} < 0
\]  

(6)

This too makes intuitive sense in the light of the fundamental trade-off described above. The principal may reduce the risk of a costly renegotiation by increasing the procured quantity, but that increases the probability of having to pay \( p \cdot q \) for a job that actually could have been carried out with less than \( q \) units of input. The higher the \( p \), the higher the cost in these outcomes. Consequently, if the expected per unit price increases, the principal will reduce the procured quantity as a renegotiation has become relatively (but not absolutely) less costly, see Figure 2d.

The optimal quantity to procure depends, in addition to the variables discussed above, on the lower and upper limit of the distribution of \( Q \). Differentiating \((2')\) with respect to \( Q_{\text{low}} \) and \( Q_{\text{high}} \) respectively yields

\[
\frac{\partial q^*_0}{\partial Q_{\text{low}}} = \frac{p}{(p+y)} > 0
\]  

(7)

\[
\frac{\partial q^*_0}{\partial Q_{\text{high}}} = \frac{y}{(p+y)} > 0
\]  

(8)

From (7) we see that if the lowest possible quantity required increases, the optimal quantity to procure must increase. As seen from (8), the same applies for the highest possible required amount. Neither of these is surprising in the light of our previous discussion. Increasing the upper boundary increases the probability of having to conduct costly renegotiations. The same applies for the lower boundary as increasing this will shift some probability mass to outcomes where \( q < Q \) and a renegotiation is required.

Equations (7) and (8) also say something about the optimal response from a change in uncertainty. For the uncertainty surrounding a given expected \( Q \) to increase it must be the
case that $Q_{low}$ decreases while $Q_{high}$ increases at the same rate. From (7) we know that the former implies a decrease in $q_0^*$, whereas, from (8), the latter implies an increase in $q_0^*$. The outcome is determined by the relative strength between (7) and (8), which depends on the relation between the expected price (being in the numerator of (7)) and the price mark-up (the numerator of (8)). We have no theoretical prediction of this. However, as long as the mark-up is less than the expected price, (7) will outweigh (8) implying that the optimal quantity to procure decreases when the uncertainty increases (also see Figure 2c).

The results above are summarized in Figure 2, which illustrates – by the means of a numerical example – the optimal quantity and resulting expected total cost as a function of the renegotiation cost, $R$, the price mark-up, $\gamma$, the quantity range, defined as $(Q_{high} - Q_{low})/2$, and the expected price, $\bar{p}$. Note that the quantity is measured on the left axis and expected cost on the right and that the scale for the expected cost is different in the expected price graph.

The expected cost, the broken lines in Figure 2, increases in $\bar{p}$, $R$ and $\gamma$ respectively. This is not surprising. All three relationships are concave\(^{12}\), which follows from that the procured quantity is calibrated to, at least partly, off-set the impact on expected cost. It is worth noting that the expected cost increases in the range around the expected quantity. That is, uncertainty is costly for the principal.

\(^{12}\) Inserting (2') in (1) and differentiating with respect to $\bar{p}$, $R$ and $\gamma$ respectively yields positive first derivatives and negative second derivatives. That is, the expected cost increases in each of the three variables, but at a decreasing rate.
Figure 2. optimal quantity to procure given no monitoring (left axis) and resulting expected total cost (right axis) as a function of $R$, $\gamma$, $(Q_{\text{high}} - Q_{\text{low}})/2$ and $\bar{p}$, respectively. The base case of this numerical example is $p_{\text{low}}=2$, $p_{\text{high}}=4$, $Q_{\text{low}}=10$, $Q_{\text{high}}=20$, $R=4$, and $\gamma=1$.

Figure 2 hints towards some other findings, in addition to those discussed above. For instance it seems as $q_0^*$ asymptotically approaches $Q_{\text{high}}$ when the price mark-up increases, and $Q_{\text{low}}$ when the expected price increases. That these observations are not artifacts of this particular numerical example is easily verified by studying (2’). There is also an intuitive explanation.

As the price mark-up becomes very high, risking a renegotiation becomes prohibitively costly. Hence, to avoid renegotiations the principal procures an amount equal (or, in the limit, very close) to $Q_{\text{high}}$. A similar logic applies for the expected price. When this is very large (relative to $R$ and $\gamma$), the principal faces strong incentives to avoid paying for more units than actually required. This is achieved by procuring an amount close to $Q_{\text{low}}$. The principal knows that this implies a large risk for a renegotiation, but the cost this incurs is relatively small.
Also note that when the range of the uncertainty surrounding $Q$ decreases, $q^*_t$ tends to the expected value of $Q$ (15 in the numerical example)\(^{13}\). This must be the case as in a setting without uncertainty; the principal clearly should procure the certain amount.

Finally, as seen from Figure 2a, when $R$ tends to zero the amount to procure is close to $Q_{low}$. This provides a good illustration of the fundamental trade-off the principal faces. The reason for a $q > Q_{low}$ when $R = 0$ in Figure 2a is that there is a positive mark-up in price in the base case of the numerical example. From (2’) it is easily seen that if both $R$ and $γ$ equal zero, the optimal quantity to procure is exactly $Q_{low}$. In the light of the discussion above, this is expected. In this setting there are no costs associated with a renegotiation and, thus, the trade-off breaks down to the corner solution of procuring $Q_{low}$, conducting the (costless) renegotiation at which the true $Q$ becomes known to both the agent and the principal. By this the principal will only have to pay for units actually required.

### 3.2 Introducing optimal monitoring

Let us now allow for positive levels of monitoring. From (3), and the assumption that $C’(α)$ is strictly positive and increasing in $α$, we may conclude that the optimal level of monitoring increases in $q$. The intuition is simple; as $q$ becomes larger, the risk for the principal of having to pay for units not actually required increases. More monitoring increases the possibilities to observe the true $Q$, which counters this risk. That is, when $q$ is large it is worth investing more in monitoring. The same logic explains why, as seen from (3), the optimal level of monitoring increases in the expected per unit price.

\(^{13}\)The range in Figure 2c starts at 2. A lower value on the range would violate the upper limit of $R$, ($(Q_{high} - Q_{low})/p)$.  

17
Intuitively, there must be a similar relationship working from $\alpha$ to $q^*$. That is, if the principal invests more in monitoring, the optimal $q$ should increase as the risk of having to pay for unnecessary units becomes less. That this intuition is correct may be seen by differentiating (2) with respect to $\alpha$. This yields

$$\frac{\partial q^*}{\partial \alpha} = \frac{p(R+(Q_{high}-Q_{low})\gamma)}{(p-\alpha+p\gamma)^2}$$

(9)

Which indeed is positive, i.e., the optimal quantity increases in the monitoring level.

To go further we need some structure on the monitoring cost function. Let us assume that $C(\alpha) = c(-\alpha - \log[1 - \alpha])$, where $c$ is a non-negative parameter. This function fulfills the requirements of $C'(\alpha) > 0$ and $C''(\alpha) > 0$. It also avoids a corner solution with full monitoring as it tends to infinity when $\alpha$ tends to 1. Using this we may derive analytical expressions for $q^*$ and $\alpha^*$ solved simultaneously. As these expressions are not easily interpretable, we illustrate the outcome with a numerical example.

Using the same values as in Figure 2, Figure 3a plots the optimal values as a function of the renegotiation cost. The solid line shows the optimal quantity. Comparing this to the no monitoring case (Figure 2a) reveals that $q^*$ now increases in a convex fashion as opposed to the linear relationship. Also, it increases faster. In the no monitoring case, $q_{0}^*$ reached its maximum value ($Q_{high} = 20$) when $R$ equaled 30. Here, $q^*$ reaches $Q_{high}$ already at $R=7.5$. This is in line with the intuition discussed above; the monitoring decreases the risk of paying for unnecessary units and thereby the cost associated with setting a high $q$.

Also in line with the intuition outlined above; the monitoring level in optimum increases in the renegotiation cost. When renegotiations are costly, the principal may initially procure
more units to avoid the risk of having to renegotiate the contract. The cost associated with this strategy may be reduced by investing more in monitoring.

Figure 3. optimal quantity to procure (left axis) and optimal monitoring level (right axis) as a function of $R$, $\gamma$, $c$ and $\bar{\rho}$, respectively. The base case of this numerical example is $p_{\text{low}}=2$, $p_{\text{high}}=4$, $Q_{\text{low}}=10$, $Q_{\text{high}}=20$, $R=4$, and $\gamma = 1$.

A similar pattern is seen for the price mark-up in Figure 3b. Just as in the no-monitoring case, the optimal quantity asymptotically tends to the upper limit as the price mark-up increases. However, it increases much faster than under no monitoring, which is due to that the expected costs of unnecessary units are held down through that monitoring also increases in $\gamma$.

Figure 3c reveals the expected relationship that optimal monitoring decreases in monitoring cost. When monitoring is reduced, so is the optimal quantity to procure. Note that when
monitoring is costless, the principal will adopt full monitoring and procure the highest possible amount, i.e., $Q_{high}$.

The optimal quantity decreases in the expected per unit price, as seen from Figure 3d. This is expected, as a higher per unit price means that the cost for any unnecessary units increases. As the quantity is reduced in optimum, generally the monitoring level is also reduced. However, for low levels of the expected price the opposite applies, i.e., then $a^*$ increases in $\bar{p}$. The reason is that the cost of monitoring is then too high relative to these low per unit prices to justify further investing in monitoring. As all the results in Figure 3, this result is specific to the cost function chosen. However, the qualitative insights apply to other cost functions fulfilling the required assumptions.

In the no-monitoring case we saw that the impact from the uncertainty, measured as the range around the expected $Q$, depends on the relative size of $p$ and $\gamma$. Figure 4 illustrates a similar, but more complex, relationship when we allow for monitoring\(^{14}\). The solid lines plot $q^*$ for three different values of the price mark-up, whereas the broken lines plot $a^*$. When the price mark-up is low (the thin lines where $\gamma=0.75$), we see that the optimal quantity is decreasing in range for all values shown in Figure 4. However, for higher values of $\gamma$, Figure 4 reveals a convex pattern where the optimal quantity decreases in range for low values, but increases when the range becomes greater. The optimal monitoring increases in the range for all values of the mark-up exhibited in the Figure. This is because a larger uncertainty implies a greater risk of having to pay for unnecessary units and monitoring is a way to reduce this risk. Also, the optimal monitoring level increases at a higher rate for larger mark-up values.

\(^{14}\)Figure 4 relies on the same numbers as Figure 3 except for $R$ being set to 1 rather than 4. This is to avoid reaching the upper limit at which $R$ exceeds the entire possible gain from setting $q < Q_{high}$. 

20
There are two forces at work simultaneously. When the uncertainty increases there is firstly a force similar to the one in the no-monitoring case. This is ultimately driven by the relative difference between per unit prices before and after a renegotiation. The second force stems from that a higher level of uncertainty calls for more monitoring in optimum. As before, when the monitoring level increases, so does the optimal quantity to procure. Due to that the monitoring level increases at a faster rate when } γ \text{ is large}, this will further pronounce the effect of increased uncertainty so that even if the optimal quantity decreases in range for low values, it may increase for higher values. Thus, we get convex patterns as in Figure 4.

This also enables us to say something about the impact of relaxing the assumption about a uniform distribution around the expected required quantity. This assumption is not very realistic and has an impact on the outcome. Without data it is difficult to say much about the shape of a realistic distribution. A starting point would be a more bell-shaped one having the same upper and lower bound as the uniform distribution. This increases the probability for outcomes close to } E\{Q\} \text{ and thus it will have a similar impact on the result as decreasing the range of the uniform distribution. As shown above the outcome of this depends on the relative
size of the per unit price and the price mark-up. The optimal monitoring level will thus decrease and, as long as the price mark-up is relatively low, $q^*$ will increase compared to the uniform distribution case.

4.0 Conclusions

We have, by the means of a simple model, studied the optimal quantity to include in the procurement document when unit price contracts are used, given that there is uncertainty around what the actual required quantity will be. The model shows that the optimal quantity to include in the procurement document, i.e., the one that minimizes the procurer’s expected total cost, is determined by a fundamental trade-off between (1) the risk of having to pay for more units than actually necessary and (2) the risk of having to conduct costly renegotiations. In optimum, the procured quantity will decrease in the expected per unit price as this reduces the costs associated with (1) and it will increase in costs associated with a possible renegotiation, which increases the costs associated with (2).

The procurer may monitor the agent during the construction phase and then receive information about how much work is conducted. This reduces (1), i.e., the risk of having to pay for unnecessary units. We have modeled this by letting the procurer invest in monitoring. A higher level of monitoring costs more, but gives a more exact (but generally not perfect) signal about the number of units actually required to finalize the project. More monitoring thus allows for procuring a larger quantity in optimum. The opposite is also true, when the procured quantity is large so is the optimal level of monitoring. Thus, the two variables must be set simultaneously in order to minimize the expected total cost.

When the procured quantity is low compared to the expected quantity, the risk that the final amount exceeds the procured one is obviously large. This implies that the actual total cost,
with high probability, will exceed the total sum calculated from the assumed quantities in the contract. If we allow ourselves to define this as being a measure of cost-overrun, this leads to an interesting conclusion. Namely that not only is it in a case like this rational and optimal to procure in a way that can be expected to lead to cost-overruns. It is actually more likely, in optimum, to see cost-overruns in projects that are expected to run smoothly in the sense that the costs of renegotiations are expected to be low. Similarly, when monitoring is costly the optimal monitoring level is lower. In optimum, the announced quantity is then also lower. Consequently, the probability for cost-overruns increases. Furthermore, we have seen that, subject to that the per unit price will not increase too much following a renegotiation, the optimal quantity to procure decreases when the uncertainty around the expected quantity increases.

That is, if one observes cost-overruns defined as above, this does not necessarily follow from any miscalculations or other errors – intentional or not – on behalf of the procurer. Rather, it may serve as an indicator that (i) the relationship between principal and agent was expected to run smoothly with low costs associated with renegotiation, (ii) that the per unit price is expected to be high relative to costs associated with a renegotiation, (iii) that the characteristics of the project is such that it is difficult, i.e., costly, to monitor the agent and/or (iv) it is difficult to make a precise ex ante estimate of the quantities required to finalize the project and the mark-up in per unit price is expected to be low. It may very well be that the procurer has correctly identified (i), (ii), (iii) and/or (iv) and responded rationally to this.

Let us conclude this article by pointing at an area for future research. The literature on unbalanced bidding predicts that a rational informed agent will exploit her superior information in the bidding process. In particular, she will post high per unit bids on activities
that she believes are underspecified by the procurer. That is, the optimal procurement strategy described above, may invoke strategic responses from the bidding agents. Understanding the implications of such strategic responses requires a unified model capable of handling both procurer and agent behavior.
References


25

