Received: 21 April 2021 | Revised: 23 September 2021 | Accepted: 9 October 2021

DOI: 10.1002/bdm.2265

RESEARCH ARTICLE

WILEY

When 2 + 2 should be 5: The summation fallacy in time prediction

Torleif Halkjelsvik[1] | **Magne Jørgensen**[1,2]

[1]Department of IT Management, Simula Metropolitan Center for Digital Engineering, Oslo, Norway

[2]Department of Computer Science, Oslo Metropolitan University, Oslo, Norway

Correspondence
Torleif Halkjelsvik, Simula Metropolitan Center for Digital Engineering, Pilestredet 52, 0167 Oslo, Norway.
Email: torleif@simula.no

Abstract

Predictions of time (e.g., work hours) are often based on the aggregation of estimates of elements (e.g., activities and subtasks). The only types of estimates that can be safely aggregated by summation are those reflecting predicted average outcomes (expected values). The sums of other types of estimates, such as bounds of confidence intervals or estimates of the mode, do not have the same interpretation as their components (e.g., the sum of the 90% upper bounds is not the appropriate 90% upper bound of the sum). The present research shows that this can be a potential source of bias in predictions of time. In Studies 1 and 2, professionals with experience in estimation provided total estimates of time that were inconsistent with their estimates of individual tasks. Study 3 shows that this inconsistency can be attributed to improper aggregation of time estimates and demonstrates how this can produce both overestimation and underestimation—and also confidence intervals that are far too wide. Study 4 suggests that the results may reflect a more general fallacy in the aggregation of probabilistic quantities. The inconsistencies and biases appear to be largely driven by a tendency to naïvely sum ($2 + 2 = 4$) probabilistic (stochastic) values. This *summation fallacy* may be consequential in contexts where informal estimation methods (expert judgment) are used.

KEYWORDS
aggregation, bias, confidence, probability, time prediction

1 | INTRODUCTION

Team members A, B, and C are quite confident that their current work tasks will require less than 50, 100 and 150 work hours, but can they be equally confident that the total amount of work will be less than 300 work hours? Your most frequent commute time is 20 min, but is your most frequent total commute time for a 5-day week 100 min? The present article describes problems relating to the aggregation of estimates and demonstrates inconsistencies and biases in people's judgments of total time usage.

People often predict that they will complete tasks earlier than they actually do (Buehler et al., 1994). When these predictions are made in spite of knowledge about past completion time, this overoptimism is referred to as “the planning fallacy” (Griffin & Buehler, 2005). This type of overoptimism is typically found for judgments of when a task will be completed, but not generally for judgments of the time required for the work itself (Halkjelsvik & Jørgensen, 2012). The former type of judgement has been referred to as completion time prediction and the latter as performance time or task duration prediction (i.e., time on task).

On average, across a range of performance time predictions in laboratory tasks in psychology, there is no general tendency of under- or overestimation. However, the amount of work in larger real-life projects is often underestimated (Halkjelsvik & Jørgensen, 2012). Studies from psychology, engineering, and management science have identified a range of factors that determine the direction of bias in performance time estimates. For instance, smaller tasks are relatively overestimated in comparison with larger tasks. This is a well-known phenomenon in research on quantitative judgments (e.g., Vierordt, 1868) and has, for example, been referred to as "the central tendency of judgment" (Hollingworth, 1910). Another factor that influences the bias is the extent of decomposition. Decomposing larger tasks into smaller ones and aggregating these estimates have been found to produce higher estimates than single estimates of the totality (e.g., Forsyth & Burt, 2008). A similar effect of higher estimates has been observed when tasks are unpacked into smaller components that are identified and described, but not separately estimated (e.g., Kruger & Evans, 2004).

There are many reasons why overall holistic predictions would differ from decomposed or unpacked estimates. For example, the abovementioned central tendency of judgments may produce overestimation (or less underestimation) as a function of the number and size of the subtasks. According to people's judgments, small subtasks are believed to require disproportionally more time than larger tasks (Halkjelsvik et al., 2011). The amount of bias also depends on which types of subtasks one is able to identify (Hadjichristidis et al., 2014). For example, if important subtasks are missing, the estimated time will be too short. In the present research, we will not compare overall holistic approaches (no decomposition) to decomposed or unpacked estimation. Instead, we will focus on the situation when estimates of elements *and* a totality are required. This includes cases of decomposition, but does not require that the starting point is a totality that needs to be broken down before estimation. Combination of elements into an estimate of the total can be meaningful without an explicit decomposition process, for example when estimating the total costs of a portfolio of projects.

1.1 | Probabilistic estimates

The same task may take 20 work hours in one case and 25 work hours in another, depending on the skills of the persons assigned to the task, their work disturbances, their choices regarding how to solve the task, and other more or less random influences. This means that the potential time needed to complete a task can be considered as a probability distribution. For example, there may be 60% chance that a task will take less than 25 work hours and a 30% chance that it will require between 20 to 25 h. Accordingly, the estimated uncertainty of a task can be expressed as probabilities using percentiles, indicated as "pX," where the X represents a specific percentile or probability (p) of an estimated outcome distribution. For example, if our p10 estimate for a task is 15 work hours and the p90 estimate is 45 work hours, we believe there is only a 10% chance of an outcome lower than 15 work hours and a 10% chance of an outcome higher than 45 work hours. We can also use these estimates to form an 80% confidence interval from 15 to 45 work hours.

Judgments of confidence intervals that include values in the farther tails of the distributions, such as estimates from p5 to p95 or 80% confidence intervals, are often too narrow (Connolly & Dean, 1997; Jørgensen et al., 2004; Jørgensen & Moløkken, 2004). When people provide too narrow confidence intervals, this is typically referred to as *overconfidence*. Overconfidence appears to be a general phenomenon that is observed across many domains (e.g., Glaser et al., 2013; Soll & Klayman, 2004; but see Gigerenzer, 2018).

The pX format can also be used to express point estimates. For example, the p50 (the prediction of the median) is commonly used as a "best guess" estimate for budgeting and planning as it gives an equal probability of overrun and underrun, and the p85 (25% chance of overrun) can be used as input to form a conservative budget (see Welde, 2017). We can also use other parameters of the probability distribution as point estimates. Common types of estimates are the "most likely" value (the outcome with the highest probability), which is the mode of potential outcomes, and the average/mean outcome, which is often referred to as the *expected value* or the *expectation*. The mean outcome can be expressed as the sum of outcomes divided by the number of outcomes if we repeated the task infinitely without learning.

The estimate type is not always clearly defined. For example, one could be interested in a cautious estimate of the total costs of renovating an apartment, a safe estimate of the time needed to run several errands (to set the parking meter), or an optimistic target to increase motivation and productivity (cf. Nan & Harter, 2009). Such verbal expressions imply that outcomes are not fully under our control and implicitly acknowledge that the final outcome is a realization of an underlying distribution of potential outcomes.

1.2 | The sum of probabilistic estimates

When aggregating estimates of multiple tasks, the mean estimates of the elements are the most useful, because the sum of the expected values of each element is equal to the expected value of the sum of the elements. This is referred to as the linearity of expectation; adding the means of random variables gives the mean (expected value E) of the total, $E[X + Y] = E[X] + E[Y]$ (see, e.g., Fristedt & Gray, 1997). Non-mean estimates do not have this property. For example, we cannot simply sum the p80 estimates for multiple elements and obtain the p80 of the total time (except when the p80s are perfectly correlated or when they happen to be equal to the mean values). To illustrate, we provide two examples. In the first example, the sum of p80 estimates does not give a 20% chance of overrun, but a less than 1% chance.

Example 1. Assume that we have perfectly calibrated p80 estimates of the time needed for 8 identical, uncorrelated tasks. For simplicity, also assume that the outcomes are drawn from symmetric normal distributions. Using the linearity properties of the variance, $Var(\sum X_i) = \sum Var(X_i)$, and of the mean, $E(\sum X_i) = \sum E(X_i)$, the correctly summed distribution has a mean of 8*E(X) and a variance of 8*Var(X). Given a standard normal distribution with $\mu = 0$ and $\sigma = 1$, the naïve sum of p80-estimates is $n^* \ F^{-1}(0.8) \approx 8^*(\mu + 0.8416^*\sigma) = 6.7328$, where the value 0.8416 comes from the inverse normal cumulative distribution function. If we look up the value 6.7328 in the cumulative distribution of the sum of the 8 distributions (which has the mean $8^*\mu = 0$ and a standard deviation of $\sqrt{(8\sigma^2)} = 2.8284$), we find that this value gives a probability of 99.14%. That is, when we estimate the p80 for our eight tasks above, the sum of the eight estimates is not the p80 for the *total* time, but approximately the p99.

A perhaps more intuitive illustration relating to Example 1 has been provided by Savage (2009, p. 70). With a wheel-of-(mis)fortune spinner, taking values from 0.0 to 0.9 (in increments of 0.1), he demonstrated how a 20% risk of financial ruin (a spin below 0.2) was greatly reduced from one spin to the average of two spins (8% risk of financial ruin). According to Savage, approximately half of his graduate students failed to see how the original uniform distribution changes to distributions with thinner tails and more observations in the middle of the scale (i.e., probabilities of high and low outcomes are reduced) when going from one to the average of multiple spins.

Example 1 suggests that naïve summation of bounds of uncertainty intervals, such as p10 and p90 estimates, respectively, can produce highly biased total estimates. Although it is possible to use estimates such as the p85 as conservative point estimates (see Welde, 2017), the most typical point estimates are those corresponding to the estimated p50 and the estimated mode. In the case of symmetric and normally distributed outcomes, as in our Example 1, using the mode and the p50 would not be a problem as they both coincide with the mean (the expected value). However, if project tasks have outcome distributions that are skewed in such a way that the mode values are lower than the mean values, the naïve sum of the mode values of project tasks is no longer the mode value in terms of the total (the project).[1] This is illustrated in the following example.

Example 2. A gamma distribution with shape parameter k and scale parameter θ, with both parameters being positive real numbers, has skewness $2/\sqrt{k}$, which is positive and implies a right-skewed distribution. The naïve sum of the mode values of *n* identically distributed and independent gamma distributions would give $n^*(k-1)^*\theta$. The shape and scale parameters of the correctly calculated sum of the distributions are n*k and θ, respectively. The correct mode value of the sum distribution is $(n^*k-1)^*\theta = n^*k^*\theta - \theta$, which is higher than the naïve sum $n^*(k-1)^*\theta = n^*k^*\theta - n^*\theta$ for all $n > 1$. The underestimation of the mode is in this case $(n^*k^*\theta - \theta) - (n^*k^*\theta - n^*\theta) = (n-1)^*\theta$.

As explained and illustrated in Examples 1 and 2 above, the summation of confidence bounds or optimistic/conservative estimates is problematic in general, whereas the summation of point estimates of the central tendency (the mode and the median) is problematic only in the case of skewed outcome distributions. There are very few publications on the shapes of performance time distributions, but the available data suggests that the distributions are typically right-skewed, with mean values higher than the medians and modes. This is illustrated in Figure 1a, which presents the distribution

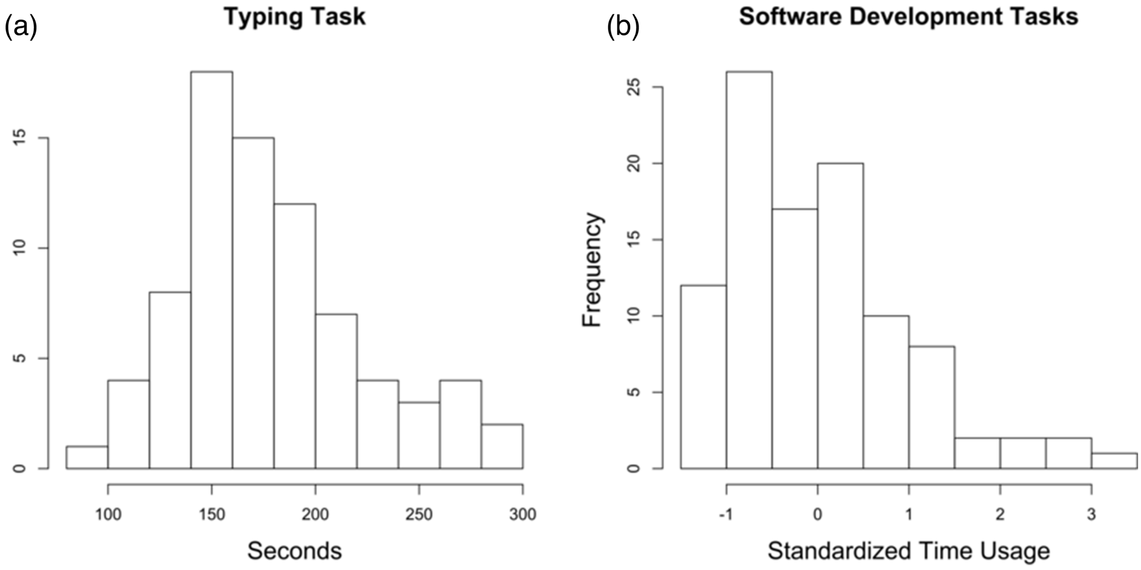

FIGURE 1 Outcome distributions of (a) seconds spent on a computer typing task and (b) work effort for five software development tasks (standardized with M = 0 and SD = 1 for each task)

of time spent on a typing task by 78 student participants in a controlled psychology experiment (Halkjelsvik, Rognaldsen, et al., 2012), and in Figure 1b, which presents the time spent on the same five development tasks by 20 experienced software developers (Jørgensen & Grusche, 2009).

There are some operations that can have normal or even left-skewed outcome distributions (e.g., Abourizk & Halpin, 1992), but in computer programming (e.g., Figure 1b), office tasks (e.g., Roy & Christenfeld, 2007, original data), or any decision time (e.g., Smith & Ratcliff, 2004) or response latency paradigm (e.g., Fazio, 1990), the outcome distributions of human operations are found to be right-skewed. Accordingly, in perhaps most estimation contexts involving time, one should not aggregate point estimates that reflect the mode outcome or the p50 (median) by simple summation, because the resulting estimate will be biased downwards.

1.3 | The present research

In Study 1, we investigated whether software professionals' estimates of the mode (most likely value) of the total time required for a set of coding tasks were too low, as inferred from estimates for the individual tasks. Study 2 is similar to Study 1 but also include assessments of confidence intervals. In Study 3, we gave participants a distribution of commute time and asked for predictions of the total time required for 10 drives. This allowed us to calculate the normatively correct predictions based on the distribution of past driving time and to test whether the participants gave biased predictions and confidence intervals. Study 4 did not concern prediction of time but instead judgments in a different quantitative domain (prediction of benefits/saved costs).

2 | STUDY 1

As described in the introduction, the outcome distributions of performance time are typically right skewed in such a way that the mode is lower than the mean value. Therefore, when requiring predictions of the mode values for a set of tasks, the estimate of the mode of the totality should be higher than the sum of the estimates of the individual tasks (see Example 2). In the present study, we calculate the sum of mode predictions for three individual tasks and compare this sum with an estimate of the total time required for all tasks. If the total estimate is equal to the sum of the estimates for individual tasks, the total estimate is too low (given that we can rely on the estimates of the individual tasks). The total estimate is also too low if it is even lower than the sum of the individual estimates.

Estimates that are higher than the sum of individual predictions can still be too low, but in the absence of data on actual performance time, we cannot know for sure. Thus, the definite threshold of a "too low" total estimate in the current context is the sum of component estimates.

2.1 | Method

2.1.1 | Sample

We recruited 57 software professionals from a software development consulting company in Poland (median years of experience = 6, range = 1 to 22). All the participants had at least half a year's experience in Java programming. Fifty-six percent had estimated software development work more than 50 times, and only 12% had estimated work less than three times.

2.1.2 | Procedure and materials

The study was carried out in the context of validating an assessment tool for programming skills and was introduced before the skills test. Participants received specifications of a Java programming task and were instructed to estimate the *most likely effort* (mode effort) they would need to complete the task. Note that in the domain of Software Development, "work effort" refers to the amount of work (i.e., "time on task"/"performance time"). Responses were entered into two boxes, one for hours and one for minutes, presented after the heading "Estimated most likely use of effort on Subtask [A/B/C]" and the text "I think I most likely will use…" This was done for three different programming tasks. Subsequently, the participants estimated the total time required for the tasks after the heading "Estimated most likely TOTAL use of effort (total most likely effort to complete Subtasks A, B and C)" and the phrase "I think I most likely will use," followed by the same response boxes as above (hours and minutes).

The software specifications and instructions were given on-screen, using the web-based survey software Qualtrics. The three programming tasks involved correcting an error in a software system

TABLE 1 Percentages (frequencies) of the samples who provided inconsistent total and component estimates and who naively added the component estimates, Studies 2 and 3

		Inconsistent			Naïve summation		
	N	Mode (total ≤ sum)	Lower bound (total ≤ sum)	Upper bound (total ≥ sum)	Mode (total = sum)	Lower bound (total ≤ sum)	Upper bound (total ≥ sum)
Study 1	57	79% (45)	NA	NA	63% (36)	NA	NA
Study 2	52	62% (32)	56% (29)	44% (23)	37% (19)	33% (17)	27% (14)

managing lab orders, extending the lab ordering process to include a commenting field, and adding a new command for delivering lab orders. The full task specifications are provided in Materials S1.

2.2 | Results and discussion

Approximately 79%, 95% CI [67, 88], of the participants provided mode estimates of the total time that were too low according to our criteria (see Table 1). The majority of these inconsistent responses reflected total estimates that were equal to the sum of the elements.

Our claim that the total estimates of the participants tended to be too low rests on the assumption that the outcome distributions of software development tasks are right-skewed, as argued in the introduction. However, it could be the case that the participants implicitly assume that outcome distributions are symmetric, which means that the mode equals the mean. This would justify a naïve summation of the elements' according to the linearity of expectation principle. To rule out this interpretation, Study 2 included questions about 90% confidence intervals.

3 | STUDY 2

Estimates that reflect the central tendency, such as the mode and the median, may be aggregated by simple summation when distributions are symmetric and unimodal, because the measures of central tendency equal the mean value. In contrast, the simple summation of other types of estimates is not meaningful even when the outcome distributions are symmetric (see, e.g., Otley & Berry, 1979). This means that the naïve addition of minimum estimates and of maximum estimates such as the p5 and p95, respectively, is problematic. As multiple extreme values are less likely than a single extreme value, the sum of lower bounds will tend to be too low to represent the true lower bound of the total, and the sum of the upper bounds will tend to be too high to represent the true upper bound of the total. Combined, this could lead to overly conservative (too wide) confidence interval for the total time.

A belief that outcome distributions of performance time are symmetric may be a valid excuse for summing point estimates such as the mode and the median, but it is not a valid excuse for naively summing elements' minimum or maximum values. Thus, in Study 2, we required confidence intervals for the set of tasks used in Study 1. By comparing the difference between the mode and the two confidence bounds, we could also assess whether the participants assumed that the distributions were symmetric.

3.1 | Methods

Fifty-two participants were recruited from the Ukrainian branch of a multi-national company (median years of experience = 6, range = 1 to 26). All the participants had at least half a year's experience in Java programming, and all the participants had experience in estimating software development work; 80% had estimated the time required for development tasks more than 50 times.

The procedure was the same as in Study 1, except for the addition of questions about 90% confidence intervals for each task and for the total. Participants received the following phrase: "I think it is about 90% likely (almost sure) that my actual use of effort on Subtask A [B/C] will be between..." and responded in hours and minutes in separate boxes for the "minimum effort" and the "maximum effort." In the same format, the participants responded to the phrase "I think it is about 90% likely (almost sure) that my total use of effort on Subtasks A, B and C (total effort) will be between..."

3.2 | Results and discussion

If we first look at the mode predictions, we observe that approximately 62%, 95% CI [48, 74], provided total estimates that were too low (see Table 1), given that our assumption of right-skewed outcome distributions hold. The majority of these represented cases where the total estimate was equal to the sum of elements.

We explored whether the participants assumed that the outcome distributions were symmetric and whether this determined the tendency to naïvely sum the estimates. Over the three estimation tasks, 44–46% of the participants provided estimates that reflected positively skewed distributions (absolute difference from the most likely estimate to the maximum was higher than to the minimum), 27–35% reflected symmetric outcome distributions, and 19–29% reflected negatively skewed distributions. Naïve summation was not related to a belief in a symmetric distribution. Among those who assumed symmetric outcome distributions, 29–33% naively summed estimates of the most likely outcome, and among those who assumed skewed distributions, this number was 38–41% (chi-squared tests, $ps > .5$ for all three tasks).

The estimates of confidence bounds revealed a substantial proportion of total estimates that were inconsistent with the component estimates. Approximately 56%, 95% CI [42, 68], of the lower bounds of the totality were too low, and 44%, 95% CI [32, 58], of the upper bounds were too high. In the majority of these cases, the inconsistency between total and component estimates was due to the naïve summation of probabilistic quantities, as inferred from the observation that the total estimate was equal to the sum of elements (see Table 1).

These result cannot be explained by a belief in symmetric and normally distributed outcomes. A belief in symmetric outcome distributions could justify the naïve summation of measures of central tendency (e.g., mode), but not summation of confidence bounds (see Example 1). A remaining interpretational issue is whether the inconsistency is due to the aggregation, the elements, or both. In principle, the aggregate estimate could be accurate and the elements biased. Records of actual performance time on the software development tasks could be useful in this respect. However, if the total estimates turned out to be more accurate than the component estimates, this could be related to characteristics of the specific tasks, rather than the aggregation process (we know from past research that these particular tasks are typically overestimated). A better way to assess if there is a bias relating to the aggregation is to provide information about the actual performance time required for the components and then assess the potential bias in the total estimate. This was the idea in the next study.

4 | STUDY 3

Studies 1 and 2 demonstrated that for a substantial proportion of the participants, estimates of the totality were inconsistent with estimates of elements. However, we were not able to determine whether the aggregation actually produced bias in the total estimates. Smaller tasks are often overestimated (e.g., Hollingworth, 1910; Vierordt, 1868), and confidence intervals are often too narrow (Connolly & Dean, 1997; Soll & Klayman, 2004). It could be the case that people's aggregation strategies compensate for their own tendency to overestimate smaller tasks. To investigate whether the aggregation process produce bias, we gave the participants in Study 3 a record of historical time usage data. Assuming that future time usage is similar to the historical record, we can compare people's predictions with sums of values drawn from the historical data to assess the level of bias.

4.1 | Method

We asked attendees at a seminar on digitalization and data management to participate in the study. Their professional work involved software architecture, large database management, application development and maintenance, user adoption, and project administration and management. Participation was voluntary and was not a required part of the seminar. Eighty-five (of approximately 90) attendees started the online survey, and 76 answered at least one of the questions pertaining to the study.

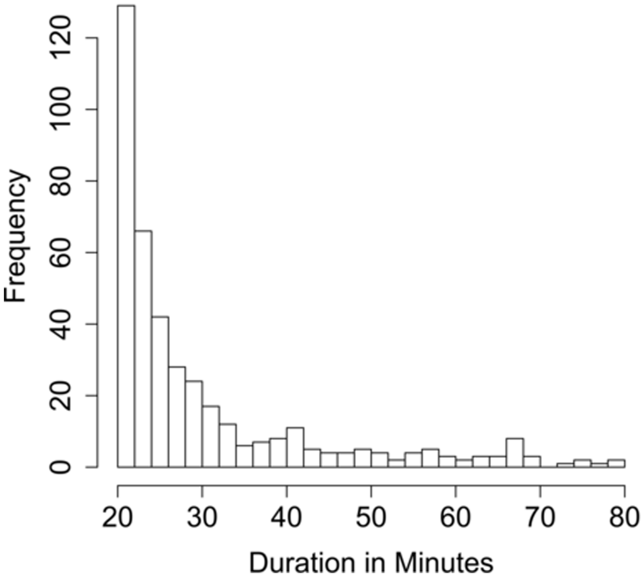

FIGURE 2 Histogram of past commute time shown to the participants in Study 3

The participants used their own cell phones, tablets, or computers to complete the online survey. On the first page they were shown a histogram with the distribution of commute time from home to work for a fictitious person called "Sivert" (see Figure 2). The 10th percentile (21 min), the 90th percentile (51 min), and the most frequent values (20–22 min) of the distribution were provided in text. The participants were first asked control questions regarding the prediction of a single drive from home to work. This allowed us to exclude inattentive participants and participants who failed to understand the concept of probabilistic estimates, together with those who failed the transition from historical data to predictions. The participants were subsequently given brief vignettes explaining that Sivert needed estimates for the most likely total time, the p10, and the p90, of the next 10 days' commute time from home to work. As an example, the request for the p90 read:

> Sivert would also like a high and conservative estimate for the sum of his driving time over the next 10 days. This estimate should be so high that he can be 90% sure that he will not spend more time than this (although there is still a 10% possibility that he will use more time than this estimate). Provide, in number of minutes, a high and conservative estimate of the SUM of his driving time to work over the next 10 workdays. Sivert can be 90% sure that the SUM of his driving time to work over the next 10 days will be lower than: [response in minutes]

The full record of vignettes and questions can be found in Materials S2. All responses were made in an open-ended format (free text, no pre-defined options).

On the last page, the participants were asked to assess the mean of the distribution shown in Figure 2 (i.e., a parameter of the observed distribution, not a prediction). The methods and the analysis plan were registered in advance of the data collection (see https://osf.io/vrq94).

4.2 | Results and discussion

In accordance with the exclusion criteria in the preregistration, we excluded data from 15 participants who failed at least one of the three control questions and from eight participants who provided answers that probably reflected predictions of single elements instead of the sum of the 10 elements (the estimates were below 100). Three single responses were omitted because they were above 1000 or below 100 min. For results on all data (no exclusions), see Materials S3.

Of the 53 participants remaining after the exclusions, 28 provided predictions of the mode that were equal to, or below, the mode values of a single drive multiplied by 10. This means that 53%, 95% CI [40, 66], provided total estimates that were too low. Of these 28 participants, 26 were defined as having naïvely summed/multiplied the mode of the historical record of single drives.

For the p10 prediction, 47 of 50 total estimates were equal to, or lower than, the naïve sum of the p10 of the elements. This means that as many as 94%, 95% CI [84, 98], provided total estimates that were too low. The majority (44 participants) were defined as having naively summed/multiplied the p10 of a single drive. For the p90 estimates, 40 of 50 predictions were equal to, or higher than, the naïve sum of the p90 of the elements. This means that 80%, 95% CI [67, 89], provided too high p90 estimates. Thirty-seven of the predictions were defined as the naïve sum of the p90 of the elements.

These were the proportions of participants giving total estimates that, without further calculations, can be defined as incompatible with the elements in the distribution in Figure 2—in a similar manner as the estimates in Studies 1 and 2. As the mode of a single drive was provided by the historical data, the empirical distribution of Figure 2 also allows us to quantify the bias.

Table 2 shows the average and the median of the judgments, in addition to tests of average bias when compared against the distribution of 100,000 simulated sums of 10 values drawn from the distribution in Figure 2. The estimates of the mode commute time for 10 days were only 13 min below the mode of the simulated sums, which was not statistically significant according to a *t* test. However, the median of these predictions was 71 min lower than the simulated true value. A Wilcoxon signed ranks test, which approximates a test of median bias, gave a *p* value of .007 (a Sign test gave $p = .001$). The large difference in the inferential statistics between the two tests was due to a few high estimates (maximum estimate = 800). When log-transforming the estimates, the *t* test of differences in means gave a *p* value of .001.

The p10 and p90 estimates showed substantial underestimation and overestimation, respectively, and gave 80% confidence intervals (the p90 estimates minus the p10 estimates) that on average corresponded to about two and a half times the correct width (i.e., 150% too wide confidence intervals). Wilcoxon and Sign tests of the median biases for the p10, the p90, and the 80% confidence interval gave *p* values below 10^{-6}. Thus, Study 3 demonstrated that the participants' aggregation strategies can produce under- and over-estimation and give confidence intervals that are far too wide. The bias was stronger in the confidence bounds than in the point estimate. This is to be expected because the point estimate represents a type of estimate that in most cases is closer in value to the mean than the confidence bounds are. Extreme outcomes are not likely to happen 10 times in a row and outcomes from opposite tails of the distribution cancel each other out. Therefore, the sum of multiple tasks converges to multiples of the mean, and the bias from naïve summation will typically be stronger for estimates representing extreme outcomes (e.g., p90) than for values representing outcomes close to the mean (e.g., the mode).

The actual mean of the distribution in Figure 2 is 30.1. When asked to assess this value directly from the empirical distribution (i.e., not for prediction), the participants did reasonably well, with a mean estimate of 28.6 (SD = 7.7). Nevertheless, more than half of the participants provided judgments below the actual mean, which may justify some of the underestimations of the mode and the p10. That is, if the mean of the distribution had been lower, the normatively correct predictions should be lower than the true values reported in Table 2, which in turn would mean that the reported bias of the p10 and mode predictions should be lower. However, in that hypothetical case, the predictions of the p90 would be even more severely biased than reported in Table 2. Inaccurate perception of the mean of the distribution can therefore not explain the bias reported in Study 3.

As an aid to understanding the histogram of the commute times, the instructions included the p10, the p90, and the mode (these values can also be read directly from the figure). Furthermore, the participants provided the p10, p90, and mode for one drive before giving estimates for multiple drives. One may therefore argue that the participants simply used the numbers given to them by the experimenter. This is a valid point, but the situation also resembles informal estimation contexts. If you have set yourself a number of cautious estimates for different parts of your renovation project, or if a project manager has received optimistic targets from employees, there is rarely information about a supplemental "mean" estimate that one can use to adjust the total estimates. Typically, people first produce or gain access to point estimates of elements (and perhaps

TABLE 2 Participants' judgments (in minutes), simulated true values, and tests of differences, Study 3

	Judgment								
	N	M (SD)	Mdn	True value	Mean bias	*t*	*p*	Mdn bias	Signed ranks *p*
Mode	53	278 (123)	220	291	−13	0.79	.22	−71	.007
p10	50	215 (41)	210	250	−35	6.00	<.001	−40	<.001
p90	50	488 (113)	510	358	+130	8.15	<.001	+152	<.001
80% P.I.	50	280 (127)	300	108	+172	9.00	<.001	+192	<.001
distr.mean	50	29 (8)	27	30	−1.5	1.38	.17[a]	−3.6	.004[a]

Abbreviations: distr.mean, judgment of the historical distribution's mean value; Mdn, median; P.I., prediction interval (confidence interval).
[a] Two-tailed (all other *p* values are one-tailed).

confidence intervals), then they must rely on this information to provide aggregate estimates.

5 | STUDY 4

Estimates of time and work are influenced by many of the same phenomena as documented in research on judgments of other types of quantities, such as anchoring effects and the central tendency of judgment (see, e.g., Halkjelsvik & Jørgensen, 2012). It is therefore reasonable to assume that a tendency to inappropriately add probabilistic quantities is a general phenomenon that would also appear in other types of judgments. Instead of providing estimates of performance time, the participants in Study 4 assessed the potential benefits of developing common core functionality of a hypothetical system used to administer a public service.

In this study, we also included an experimental condition aimed at debiasing the aggregation of elements. In the debiasing condition, we stripped the task for contextual information and illustrated how the aggregation of benefits could be considered as a random draw of balls. For a range of probabilistic reasoning problems, the conversion of probabilities to frequencies can help people make better judgments (e.g., Gigerenzer & Hoffrage, 1995). In the debiasing condition, the probabilities of different outcomes were represented by the frequencies of different types of balls in a bucket of 100 balls.

5.1 | Method

5.1.1 | Sample

We recruited participants from a seminar in Oslo, Norway, on the management of software projects. Of the about 100 participants at the seminar, 66 completed the study. The participants were software managers from the public sector (about 60% of the participants) and software managers and developers from the private sector (about 40% of the participants).

5.1.2 | Procedure

We introduced the study as an estimation challenge with two prizes (approximately €100) for the two best estimators. The study was administered as a Qualtrics survey, and participants used their own computers, tablets, or phones to complete the study. Participants were introduced to the background for developing a common system to administer a "Youth Card" for subsidizing leisure activities. The vignette stated that the Norwegian Directorate for Children, Youth and Family Affairs considered developing a common core functionality to administer this card, and that this had the potential to save costs in comparison with individual administration systems within the municipalities. The potential benefits had been assessed, and the participants received a list on the following format that stated the probabilities of costs saved in NOK million (NOKm) for the average municipality: "X% likely that Y million will be saved." The following probabilities were listed: 10% for 0 cost saved; 20% for 1 million; 15% for 2 million; 10% for 3, 4, and 5 million, respectively; 5% for 6, 7, 8, 9, and 10 million, respectively. See Materials S4 for more details.

Participants were informed that, as part of a cost–benefit analysis, they were asked to estimate the most likely cost saved for a few pessimistic scenarios regarding the number of municipalities that would use the core functionality. Specifically, they were asked to provide the most likely sum of saved costs if 1, 2, and 10 municipalities were to adopt the system. They were asked to assume no learning between municipalities and that the list of probabilities applied to all municipalities. This was the *Saved Costs* condition.

As an attempt to debias the addition of probabilistic quantities, we provided another version of the experiment to half of the participants (random allocation), where almost all contextual information was removed. This constituted the *Random Draw* condition. The instructions stated: "Assume that we have multiple IT-projects that all have the same probabilities of benefits (saved costs in NOKm)." They were shown the same list as above and explained that this could be considered as a bucket of 100 balls, with the number of balls determined by the probability of the given outcome, such that there would be 10 balls with the number 0, since it is 10% likely that saved costs are 0, and 20 balls with the number 1, since it is 20% likely that 1 million is saved. The participants received an explanation of a procedure where a random ball is drawn, its number is recorded, the ball is replaced in the bucket, and another ball is drawn, etc. A list with the number of balls per outcome, instead of probabilities, was presented next to a picture of a transparent bag of numbered balls. Participants were asked: "What is the most likely benefit (in millions) for one project (most likely number obtained by a random draw of 1 ball)?" and "What is the most likely benefit (in millions) for two projects (most likely sum of the numbers obtained by a random draw of two balls, with replacement for each draw)?" An identical question to the latter was also asked for the outcome of the sum of 10 balls. All answers in both conditions were given in a free response (textbox) format.

5.2 | Results and discussion

One participant provided answers in percentages, which could not be used for analyses. Two participant provided answers in both percentages and NOKm, where we used the latter for analyses (results were not affected by excluding these two participants). This gave 65 participants, 36 in the Save Costs condition and 29 in the Random Draw condition.

Across the two conditions, 27 of 65 participants naively summed the most likely value for two projects (answered 2 NOKm), and 24 naively summed 10 projects (answered 10 NOKm). If we also count the participants who provided total estimates below the naive sum of elements, 32 (49%, 95% CI [37, 62]) and 37 (57%, CI [44, 69]) participants provided total estimates that were too low. As observed

TABLE 3 Participants' judgments (in NOKm), simulated true values, and tests of differences between judgments and true values, Study 4

	Judgment							
Condition	*N*	M (SD)	Mdn	True value	Mean bias	*t* test (*p*)	Mdn bias	Signed ranks *p*
Save costs								
One project	36	1.9 (1.3)	1	1	+0.9	4.13 (<.001)	0	<.001[a]
Two projects	36	3.9 (2.2)	3	5	−1.1	−3.06 (.004)	−2	.002
Ten projects	36	19.8 (13.8)	20	36	−16.2	−7.02 (<.001)	−16	<.001
Random draw								
One project	29	1.8 (1.3)	1	1	+0.8	3.38 (.002)	0	.005[a]
Two projects	29	3.6 (2.5)	2	5	−1.4	−3.10 (.004)	−3	.008
Ten projects	29	16.0 (12.8)	10	36	−20.0	−8.43 (<.001)	−26	<.001

Abbreviation: Mdn, median.

[a] Due to a high number of ties, this does not approximate a test of median bias, but instead tests the asymmetry for all ≠ 0 values.

in Table 3, this resulted in severe underestimation of total costs, even for the sum of two projects. For 10 projects, the total estimate was less than half the true value.

As to the debiasing condition, Table 3 shows that there was no increase in accuracy when the process was described as random draws from a bucket of balls. A test of differences between conditions gave $t(57.6) = -0.6$, $p = .53$, for the summation of two projects and $t(61.4) = 0.2$, $p = .84$ for the summation of 10 projects.

On the question about the mode value of one project, 34 answered the correct answer "1," and 19 gave an answer between 2 and 4, which was in the middle of the distribution (the answers "0" and "2" were given by four participants each; "4" and "5" were given by two participants). This could indicate that a large proportion of the participants perceived the median, mean, or the middle of the scale as the most likely outcome of a single project. Among the 19 participants who answered a middle value (between 2 and 4) on the question about the "most likely" outcome of a single project, the bias was reversed for the sum of two projects, mean bias = 1.1, $t(18) = 3.9$, $p = .001$, and attenuated but still negative for 10 projects, mean bias = −5.6, $t(18) = -2.3$, $p = .03$. Ironically, these participants gave less biased responses on the summation of 10 projects because they failed the task (they were unable to identify the mode value of the elements), which suggests that the bias can be even stronger than reported in Table 3 when the mode/most likely value is correctly identified.

6 | GENERAL DISCUSSION

An estimate of the time required for a piece of work can take very different interpretations. For example, it can represent an optimistic target, a cautious guess, or the average if we were to repeat the task. We can consider an estimate as a reference to a point or a parameter of a distribution of potential outcomes. When estimating the totality (project/portfolio) of a set of component tasks, it is important to consider which point or parameter the estimates are intended to represent. Only estimates of the mean of the outcome distribution can be aggregated by simple summation while retaining its interpretation as a mean estimate. The sum of non-mean estimates does not have the same interpretation as its components. In the present article, we showed that this statistical property can produce bias in people's predictions. The lower bounds of participants' confidence intervals for the total time usage of a set of tasks were too low and their upper bounds were too high. Due to the right-skewness of potential outcomes, total estimates of the mode were too low. The bias appeared to be largely due to the naïve summation of estimates. This summation fallacy is likely not restricted to the domain of time prediction, as the participants in Study 4 made similar errors when predicting the benefits of a project.

According to Griffin and Buehler (2005), the planning fallacy refers to a particular type of underestimation wherein a person is overly optimistic for a specific task while holding a more realistic belief in general (e.g., based on past experience or observations of others). In other words, the planning fallacy is an example of a situation characterized by the tendency to underweigh or ignore distributional information (the outside view) and overweight case-based, singular information (the inside view; cf. Kahneman & Lovallo, 1993), such as the steps involved in performing the task. Like the planning fallacy, the present summation fallacy can also give rise to underestimation. However, it is interesting to note that the fallacies are nearly opposites in terms of the underlying mechanism. Instead of ignoring distributional information, as in the planning fallacy, the participants relied excessively on the distributional information in the present studies. At least in Study 3, it is highly unlikely that the participants considered aspects of how the process of driving would unfold. That is, they did not rely on a singular, inside view, instead, they directly used the probabilities and quantities provided in the experiment.

When the participants reported the sum of p90 estimates as an p90 estimate of the totality, they did not take into account that the probabilities change from single events to the total. This is reminiscent of behavior in studies on multi-event probabilities. For example, Bar-Hillel (1973) found that people overestimate the likelihood of conjunctive events (e.g., chance that one would draw a colored marble three times when each draw had a probability of 0.5) and underestimate the likelihood of disjunctive events (e.g., chance of drawing at least one colored marble). She concluded that the probabilities of the individual elements had a greater influence on the judgments than the number of events.

A similar conclusion may be drawn for the present results as well. People rely on the probabilities of the elements and fail to acknowledge that the probability of the sum changes as a function of the number of elements. However, the combination of probabilities and quantitative outcomes (e.g., work hours) constitute an even more complex problem than typical cases of compound probabilities. For example, if the p80 estimate of three similar tasks is 60 work hours each, the p80 for the total cannot be calculated as the probability that *all* three events will be below 60 (which would be 0.8*0.8*0.8 = 51%). One way to view this complexity is to say that outcomes that exceed the estimate (e.g., 70 work hours, a 10-h overrun) can be compensated for by values lower than the estimate (e.g., two tasks of 55 work hours each). It is highly unlikely that people are able to intuitively calculate probabilistic information of this kind. The fact that the authors resorted to Monte Carlo simulations to obtain the normatively correct estimates in Studies 3 and 4 is illustrative of the demands placed on the participants. Thus, it is likely that participants reduce the problem to a much simpler one, such as by relying on the probabilities of the elements (see also Gneezy, 1996), believing that a probability of 80% also applies to the aggregate level.

Although one can frame the summation fallacy as a bias in probability judgments, one can also consider the problem as ignorance of the probabilistic nature of certain types of quantities. If one car costs €40,000, we know that two cars of the same type will cost €80,000. It is not easy to realize that the sum of a quantity can be different from the simple sum of its elements.

6.1 | Debiasing

In some statistical reasoning tasks, people understand frequencies better than probabilities (e.g., Gigerenzer & Hoffrage, 1995). Study 4 included an experimental condition in which participants received distributional information as both probabilities and frequencies. As a visual aid to underline the randomness in multiple draws, we showed a bucket of numbered balls, which represented the outcomes. The estimates of the mode of the total outcome in this experimental condition were equally biased as the estimates in the group that received no such information. Thinking in terms of frequencies may be just as difficult as thinking in terms of probabilities in the present context. It is somewhat counterintuitive that two times the value on the most frequent ball in a bucket is not the most likely (nor the most frequent) sum when drawing two balls from the bucket.

It should be noted that the participants in all our studies provided judgements without any feedback on accuracy. In the context of performance time predictions, training has not been very successful in improving estimation (e.g., Abrahamsson & Kautz, 2002; Prechelt & Unger, 2001), but for confidence assessments, training in the form of feedback may increase realism (Jørgensen & Teigen, 2002). It is reasonable to expect that a person with a task like the one given in Study 3 would update his/her assessment of the mode of 10 days' commute time after repeatedly experiencing the total outcome. It is however questionable whether this would lead to improvements in other contexts that involve aggregation of stochastic variables. Furthermore, in professional and everyday time predictions contexts, feedback is typically infrequent and delayed (typically after the work is completed). This means that the context is far from optimal in terms of developing intuitive judgment (cf. Kahneman & Klein, 2009).

The present work concerned the everyday and more intuitive types of predictions. In more formal estimation contexts, there are tools and models that can be used to derive estimates of the mean value and the estimated probability distribution. As input in the calculation of means and distributions, these methods can take estimates of the minimum, mode, and maximum outcome values; estimates of two or three percentiles of a chosen distribution; or past estimation error (e.g., Abourizkm et al., 1991; Halkjelsvik & Jørgensen, 2018, pp. 27–28; Keefer & Verdini, 1993; Mohan et al., 2007; Morris et al., 2014). In situations that typically involves aggregation by subjective judgment or simple spreadsheet calculations, the use of the tools cited above may increase the realism of total estimates, while also allowing people to provide non-mean estimates such as challenging optimistic targets that can motivate performance. The methods above are not without flaws. Most important, they require accurate assessments of uncertainty, whereas people tend to provide intervals that are too narrow in comparison with actual outcome distributions (i.e., overconfidence; Budescu, 2007; Connolly & Dean, 1997; Soll & Klayman, 2004).

Past studies' finding that people often provide too narrow confidence intervals prompts the question of whether the summation fallacy can reduce overconfidence. One reason for providing too narrow confidence intervals is that they appear more informative (Yaniv & Foster, 1995). For example, in one study, managers believed that developers who provide narrow intervals are more skilled and have more knowledge about the task than developers who provide wider (and likely more realistic) intervals (Jørgensen et al., 2004). In the present context, the desire to be accurate and informative would need to compete with the tendency to rely on the estimates of elements. One may therefore expect that incorrect aggregation of confidence intervals will reduce bias in some contexts. The problem is that accurate input at the level of elements, such as when one relies on records of past work, can produce severely biased estimates in the opposite direction (i.e., underconfidence), as demonstrated in Study 3. Naïve summation, therefore, appears as a highly unreliable remedy for overconfidence.

6.2 | The summation fallacy as a more general problem

The findings in Study 4, where participants were requested to sum potential benefits of a project, conceptually replicate those of the three preceding studies and suggest that the summation fallacy is not limited to judgments of time. In addition, we have found several anecdotal examples of the neglect of the probabilistic nature of predictions in other domains. Both the Norwegian government and the Norwegian Public Roads Administration announced as an achievement that the total costs of their respective project portfolios were below the sum of p85 estimates (Norwegian Government, 2019; Sandvin, 2016). Like many of the participants in the present studies, they simply summed the p85 predictions, and considered this sum as a reasonable cost control goal. However, the meaning of this goal changes as a function of the number of projects in the portfolio. A sufficiently large portfolio will make the sum of p85 estimates extremely difficult to overrun, even if the p85 estimates were severely biased.

The allocation letter from the Norwegian government to the Directorate of Public Construction and Property states that the total cost of the portfolio of completed projects the last 5 years should not exceed the sum of the p50 estimates (Ministry of Local Government and Modernization, 2018). Again, this is an example of how non-mean estimates are simply summed. If the outcome distributions of the construction projects are positively skewed, this is a requirement to provide p50 estimates that are not really p50 estimates. Similarly, Emhjellen et al. (2002) give the example of a large (€300 mill+) North Sea oil project where estimates of capital expenditures are given as p50 and simply summed to obtain total capital expenditures.

The above anecdotes indicate that misunderstandings relating to sums of probabilistic values can be relevant in other contexts than time. Future studies may discover new contexts where the summation fallacy can be of applied relevance (e.g., sports results, medical decision making, and household economy).

6.3 | Limitations

The present study used convenience samples, mainly of software developers, which may limit the generalizability of our findings. However, the samples in Studies 3 and 4 included a range of professionals in various roles, including managers and other administrative personnel. Furthermore, the findings were consistent across samples from low-cost (Studies 1 and 2) and high-cost countries (Studies 3 and 4), and across very different types of outcomes (software development effort, commute time, and benefits).

One important limitation is that the present studies assumed independence of tasks. In many projects the costs of the different elements are positively correlated. The naïve summation of estimates may produce less bias when there are strong positive correlations between the elements. For example, in the case of perfect correlation between the outcomes of multiple tasks, the sum of p90 estimates should be the same as the p90 estimate for the total (see Otley & Berry, 1979). Also weaker dependencies between elements can greatly reduce the bias when aggregating non-mean estimates, particularly when the type of estimate represents a value close to the mean (see Skerratt, 1982). In theory, participants may have believed that there were dependencies between tasks in Studies 1 and 2, but in Study 3, this would not be a reasonable assumption (one drive would typically not affect the next). In Study 4, we explicitly asked the participants to assume independence between outcomes.

7 | CONCLUSIONS

A substantial proportion of people with experience in estimation and project management provided total estimates that were statistically inconsistent with their estimates of individual tasks. This inconsistency was also found in estimates of the sum of elements from records of past performance time, suggesting a problem in the process of aggregation of time estimates. Inappropriate aggregation strategies can produce severe underestimation or overestimation, as well as underconfidence (i.e., confidence intervals that are too wide). The bias in aggregate estimates is largely attributable to naïve summation ($2+2=4$). By naively summing stochastic quantities as if they were deterministic, people commit what we refer to as the summation fallacy. When the estimates of elements represent values that are typically higher than the mean (e.g., p90 estimates), the total estimate based on the sum of these elements is biased upwards (overestimation), and when estimates of elements represent values that are typically lower than the mean (e.g., the mode value in right-skewed distributions), the total estimate based on the sum of elements is biased downwards (underestimation). The fallacy can be consequential in everyday estimation contexts (e.g., the sum of optimistic predictions for three different tasks is far more optimistic than the predictions of each of the tasks), in industries that rely on informal estimation methods, and as suggested by anecdotal evidence, in politicians' and other stakeholders' evaluations of portfolios of projects.

ACKNOWLEDGEMENT

The research received no particular funding.

AUTHOR CONTRIBUTIONS

Torleif Halkjelsvik: Conceptualization; data curation; formal analysis; visualization; writing – original draft; writing – review and editing. **Magne Jørgensen**: Conceptualization; data curation; investigation; writing – review and editing.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Torleif Halkjelsvik https://orcid.org/0000-0003-3851-6996

Magne Jørgensen https://orcid.org/0000-0001-6250-9783

ENDNOTE

[1] Note that a distribution with positive skew, as defined by the third moment coefficient, can in some cases (e.g., multimodal and discrete distributions) have a mean that is lower than the median and the most likely value. For some of the other measures of skewness (e.g., non-parametric skew), the mode and/or the median are per definition lower than the mean in a positively skewed unimodal distribution (see David & Seward, 2011).

REFERENCES

Abourizk, S. M., & Halpin, D. V. (1992). Statistical properties of construction duration data. *Journal of Construction Engineering and Management*, *118*(3), 525–544. https://doi.org/10.1061/(ASCE)0733-9364(1992)118:3(525)

Abourizkm, S. M., Halpin, D. W., & Wilson, J. R. (1991). Visual interactive fitting of beta distributions. *Journal of Construction Engineering and Management*, *117*(4), 589–605. https://doi.org/10.1061/(ASCE)0733-9364(1991)117:4(589)

Abrahamsson, P., & Kautz, K. (2002). Personal software process: Classroom experiences from Finland. *In European conference on Software Quality* (pp. 175–185). Berlin, Heidelberg: Springer.

Bar-Hillel, M. (1973). On the subjective probability of compound events. *Organizational Behavior and Human Performance*, *9*(3), 396–406. https://doi.org/10.1016/0030-5073(73)90061-5

Budescu, D. U. (2007). Coherence and consistency of investors' probability judgments. *Management Science*, *53*(11), 1731–1744. https://doi.org/10.1287/mnsc.1070.0727

Buehler, R., Griffin, D., & Ross, M. (1994). Exploring the "planning fallacy", why people underestimate their task completion times. *Journal of Personality and Social Psychology*, *67*(3), 366–381. https://doi.org/10.1037/0022-3514.67.3.366

Connolly, T., & Dean, D. (1997). Decomposed versus holistic estimates of effort required for software writing tasks. *Management Science*, *43*(7), 1029–1045. https://doi.org/10.1287/mnsc.43.7.1029

David, P. D., & Seward, L. E. (2011). Measuring skewness, a forgotten statistic? *Journal of Statistics Education*, *19*(2), 1–18. https://doi.org/10.1080/10691898.2011.11889611

Emhjellen, K., Emhjellen, M., & Osmundsen, P. (2002). Investment cost estimates and investment decisions. *Energy Policy*, *30*(2), 91–96. https://doi.org/10.1016/S0301-4215(01)00065-9

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick & M. S. Clark (Eds.), *Research methods in personality and social psychology. Review of personality and social psychology* (Vol. 11, pp. 74–97). Sage.

Forsyth, D. K., & Burt, C. D. B. (2008). Allocating time to future tasks: The effect of task segmentation on planning fallacy bias. *Memory & Cognition*, *36*(4), 791–798. https://doi.org/10.3758/mc.36.4.791

Fristedt, B., & Gray, L. (1997). *A modern approach to probability theory*. Springer Science+Business Media. https://doi.org/10.1007/978-1-4899-2837-5

Gigerenzer, G. (2018). The bias bias in behavioral economics. *Review of Behavioral Economics*, *5*, 303–336. https://doi.org/10.1561/105.00000092

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction, frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Glaser, M., Langer, T., & Weber, M. (2013). True overconfidence in interval estimates, evidence based on a new measure of miscalibration. *Journal of Behavioral Decision Making*, *26*, 405–417. https://doi.org/10.1002/bdm.1773

Gneezy, U. (1996). Probability judgments in multi-stage problems: Experimental evidence of systematic biases. *Acta Psychologica*, *93*(1–3), 59–68. https://doi.org/10.1016/0001-6918(96)00020-0

Griffin, D., & Buehler, R. (2005). Biases and fallacies, memories and predictions: Comment on Roy, Christenfeld, and McKenzie (2005). *Psychological Bulletin*, *131*(5), 757–760. https://doi.org/10.1037/0033-2909.131.5.757

Hadjichristidis, C., Summers, B., & Thomas, K. (2014). Unpacking estimates of task duration: The role of typicality and temporality. *Journal of Experimental Social Psychology*, *51*, 45–50. https://doi.org/10.1016/j.jesp.2013.10.009

Halkjelsvik, T., & Jørgensen, M. (2012). From origami to software development: A review of studies on judgment-based predictions of performance time. *Psychological Bulletin*, *138*(2), 238–271. https://doi.org/10.1037/a0025996

Halkjelsvik, T., & Jørgensen, M. (2018). *Time predictions, understanding and avoiding unrealism in project planning and everyday life*. Springer Open.

Halkjelsvik, T., Jørgensen, M., & Teigen, K. H. (2011). To read two pages, I need 5 minutes, but give me 5 minutes and I will read four: How to change productivity estimates by inverting the question. *Applied Cognitive Psychology*, *25*(2), 314–323. https://doi.org/10.1002/acp.1693

Halkjelsvik, T., Rognaldsen, M., & Teigen, K. H. (2012). Desire for control and optimistic time predictions. *Scandinavian Journal of Psychology*, *53*(6), 499–505. https://doi.org/10.1111/j.1467-9450.2012.00973.x

Hollingworth, H. L. (1910). The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, *7*(17), 461–469. https://doi.org/10.2307/2012819

Jørgensen, M., & Gruschke, T. M. (2009). The Impact of Lessons-Learned Sessions on Effort Estimation and Uncertainty Assessments. *IEEE Transactions on Software Engineering*, *35*(3), 368–383. https://doi.org/10.1109/tse.2009.2

Jørgensen, M., & Moløkken, K. (2004). Eliminating over-confidence in software development effort estimates. In F. Bomarius & H. Iida (Eds.), *Product focused software process improvement* (Vol. 3009). PROFES 2004. Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/978-3-540-24659-6_13

Jørgensen, M., & Teigen, K. H. (2002). *Uncertainty intervals versus interval uncertainty: An alternative method for eliciting effort prediction intervals in software development projects* (pp. 343–352). Singapore: In International conference on project management (ProMAC).

Jørgensen, M., Teigen, K. H., & Moløkken, K. (2004). Better sure than safe? Over-confidence in judgement-based software development effort prediction intervals. *Journal of Systems and Software*, *70*(1–2), 9–93. https://doi.org/10.1016/S0164-1212(02)00160-7

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Kahneman, D., & Lovallo, D. (1993). Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking. *Management Science*, *39*(1), 17–31. https://doi.org/10.1287/mnsc.39.1.17

Keefer, D. L., & Verdini, W. A. (1993). Better estimation of PERT activity time parameters. *Management Science*, *39*(9), 1086–1091. https://doi.org/10.1287/mnsc.39.9.1086

Kruger, J., & Evans, M. (2004). If you don't want to be late, enumerate: Unpacking reduces the planning fallacy. *Journal of Experimental Social Psychology*, *40*(5), 586–598. https://doi.org/10.1016/j.jesp.2003.11.001

Ministry of Local Government and Modernisation. (2018). Letter of allocation 2019—Norwegian Directorate of Public Construction and Property. Letter to Norwegian Directorate of Public Construction and Property.

Mohan, S., Gopalakrishnan, M., Balasubramanian, H., & Chandrashekar, A. (2007). A lognormal approximation of activity duration in PERT using two time estimates. *Journal of the Operational Research Society*, *58*(6), 827–831. https://doi.org/10.1057/palgrave.jors.2602204

Morris, D. E., Oakley, J. E., & Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling and Software*, *52*, 1–4. https://doi.org/10.1016/j.envsoft.2013.10.010

Nan, N., & Harter, D. E. (2009). Impact of budget and schedule pressure on software development cycle time and effort. *IEEE Transactions on Software Engingeering*, *35*(5), 624–637. https://doi.org/10.1109/TSE.2009.18

Norwegian Government. (2019). Hva er KS-ordningen? [What is the QA arrangement?]. https://www.regjeringen.no/no/aktuelt/Hva-er-KS-ordningen/id2001422

Otley, D. T., & Berry, A. (1979). Risk distribution in the budgetary process. In C. Emmanuel, D. Otley, & E. Merchant (Eds.), *Readings in accounting for management control* (pp. 266–283). Springer Science+Business Media. https://doi.org/10.1007/978-1-4899-7138-8_13

Prechelt, L., & Unger, B. (2001). An experiment measuring the effects of personal software process (PSP) training. *IEEE Transactions on Software Engineering*, *27*(5), 465–472. https://doi.org/10.1109/32.922716

Roy, M. M., & Christenfeld, N. J. S. (2007). Bias in memory predicts bias in estimation of future task duration. *Memory & Cognition*, *35*, 557–564. https://doi.org/10.3758/BF03193294

Sandvin, B. (2016). Vi holder kostnadsrammen!. https://www.tu.no/artikler/vi-holder-kostnadsrammen/366329

Savage, S. (2009). *The flaw of averages: Why we underestimate risk in the face of uncertainty*. John Wiley & Sons.

Skerratt, L. C. L. (1982). Risk distribution in the budgetary process, a comment. *Accounting and Business Research*, *12*(47), 233–235. https://doi.org/10.1080/00014788.1982.9728812

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. https://doi.org/10.1016/j.tins.2004.01.006

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*(2), 299–314. https://doi.org/10.1037/0278-7393.30.2.299

Vierordt, K. (1868). *Der zeitsinn nach versuchen*. H. Laupp.

Welde, M. (2017). Cost performance in large government investment projects that have been subjected to external quality assurance. Concept report no. 51. Ex ante Academic Publisher.

Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General*, *124*(4), 424–432. https://doi.org/10.1037/0096-3445.124.4.424

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Halkjelsvik, T., & Jørgensen, M. (2022). When 2 + 2 should be 5: The summation fallacy in time prediction. *Journal of Behavioral Decision Making*, *35*(3), e2265. https://doi.org/10.1002/bdm.2265