THE
BEHAVIOURAL
INSIGHTS
TEAM

Improving NHS Procurement efficiency using digital platforms

Kristina Londakova, Abigail Mottershaw, Victoria Fussey, James Farrington, Janna Ter Meer and Hugo Harper

February 2020

Table of Contents

Table of Contents 2

Executive summary 3

1.Background 7

2. Intervention design 8

3. Trial design and implementation 16

3.1 Experiment design 16

3.2 Outcome measures 19

3.3 Sample 20

4. Results 21

4.1 Tackling spend accuracy 22

4.5 Tackling efficiency 27

4.6 Tackling satisfaction 27

5. Limitations 29

6. Implications 32

7. Recommendations 33

8. Appendices 34

9. Endnotes 44

Executive summary

Background

Up to £700 million could be saved in the NHS every year through more efficient procurement of goods and services, according to Lord Carter's review. Procurement inefficiency stems in part from substantial variations in prices paid for the same items by different NHS Trusts (for example, one hospital may pay £12 for a box of syringes and another may pay only £4) but also within the same Trust (e.g. a requisitioner may buy a pack of surgical gloves for 50p or £1.27). Whilst the first relates to Trusts' negotiation skills and coordination, the second relates to the behaviour of individual hospital requisitioners which it may be easier to influence. Qualitative research suggests that some NHS requisitioners struggle to procure products efficiently which can in turn result in expensive orders, errors or over-ordering.

In 2016, the Behavioural Insights Team (BIT) won an award from the Health Foundation to test whether changes to the choice architecture of digital procurement platforms could improve requisitioner decision-making and generate savings for the NHS. Whilst waiting for the roll-out of the new procurement platform 'Edge for Health' - developed by Virtualstock and due to be introduced in over 60 NHS Trusts - we used our online experimental platform Predictiv to simulate and test our intervention ideas.

The trial

We ran a two-arm online randomised controlled trial, working with Virtualstock, to test whether a behaviourally-informed and data-driven platform design can help requisitioners to save the NHS money by finding the cheapest options. In December 2019, we recruited almost 1,000 NHS staff to take part in this trial (65% of which made procurement decisions as part of their role). Participants completed two rounds of a hypothetical shopping exercise using BIT's simulated version of the 'Edge for Health' procurement platform. Whilst the participants in the control group completed the exercise using the standard platform, participants in the treatment group used our behaviourally-informed version. We designed this new version to target five key issues on the platform using a bundled intervention with five solutions, distributed across the whole shopping cycle (see the summary table below).

Table 1. Summary table of issues and behavioural solutions

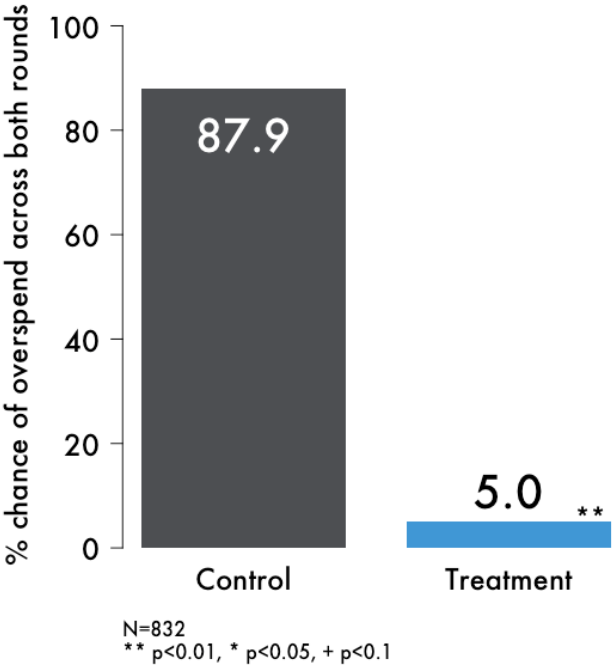
Issue	Solution
Complex product prices	**Increasing salience of cheaper alternatives**
	• Price-based ordering by cost per individual unit. By default, this ordering included the cheapest delivery costs (with an option to exclude delivery, if preferred by users); • Vertical product display to prime users to focus on the top and cheapest option; • Clear display of price per individual unit; • Green highlighting and a salient tag for the cheapest supplier, defined as the supplier who sells the product for the lowest combined cost per unit of the product and the delivery charge.
Unclear delivery options	**Defaults and alerts to optimise delivery**
	• Default to the cheapest delivery option for all products and suppliers; • Earlier delivery choice enabled at the point a product is added to the cart; • Salient alerts when a more expensive delivery option is selected at the cart stage and at the checkout, instead of a single pop up at the checkout.
Confusing quantities	**Nudge to check product quantity**
	• Clear information on different package sizes and the number of individual items throughout the platform; • Primer for quantity check at the cart, encouraging users to take time, check their ordered quantities and actively confirm they have selected the correct quantity.
Hidden savings	**Order swap**
	• An ‘order swap’ suggestion with cost feedback and suggested swaps for the given basket, using loss-framing to highlight potential missed savings to the NHS. This order swap was based on an algorithm calculating the most optimally priced basket, based on the user’s selection of products, quantities and delivery dates, taking into account potential savings on large quantities and multiple item orders.
Lack of monitoring	**Cost feedback**
	• Cost feedback on how much the requisitioner has spent on their order compared to the cheapest available basket for the identical order. To increase the sense of social accountability, we also included a monitoring message to highlight that the requisitioner’s behaviour is being

monitored and the order may be reviewed by a superior.

Results

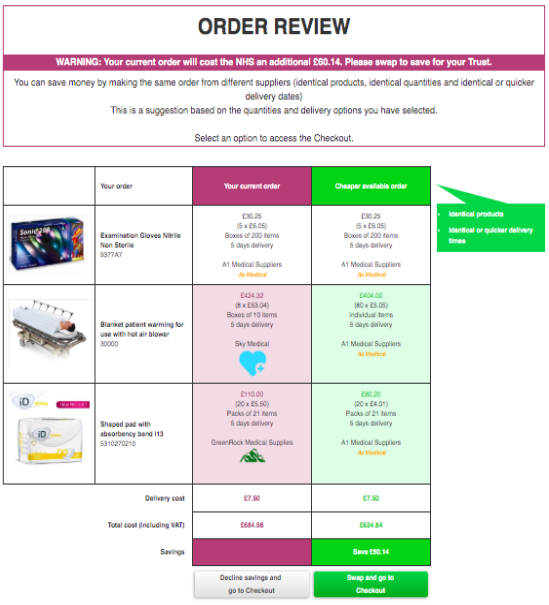
We found that our behaviourally-informed procurement platform design was effective at reducing the chance of overspend. In the treatment group, the chance of overspend was reduced by 83 percentage points compared to the control group shopping on the current platform. This effect was driven by reducing the likelihood of overspend occuring at all, rather than reducing the magnitude of the overspend.

Figure 1. Percentage change of overspend



Our exploratory analysis shows that this sizable effect was largely driven by the 'order swap' element of our intervention, the tailored suggestion of the cheapest possible basket based on the participant's chosen products, quantities and delivery. The earlier behaviourally-informed cost prompts would have actually increased overspend if used in isolation. This is because the earlier prompts flagged cheaper products per item, including the cheapest delivery, but did not account for the specific quantities or combinations of several goods that requisitioners may select. The key takeaway here is that finding the cheapest overall basket is a task of such substantial computational complexity that only a sophisticated analytical prompt can help requisitioners to realise savings.

Figure 7. Illustration of order swap



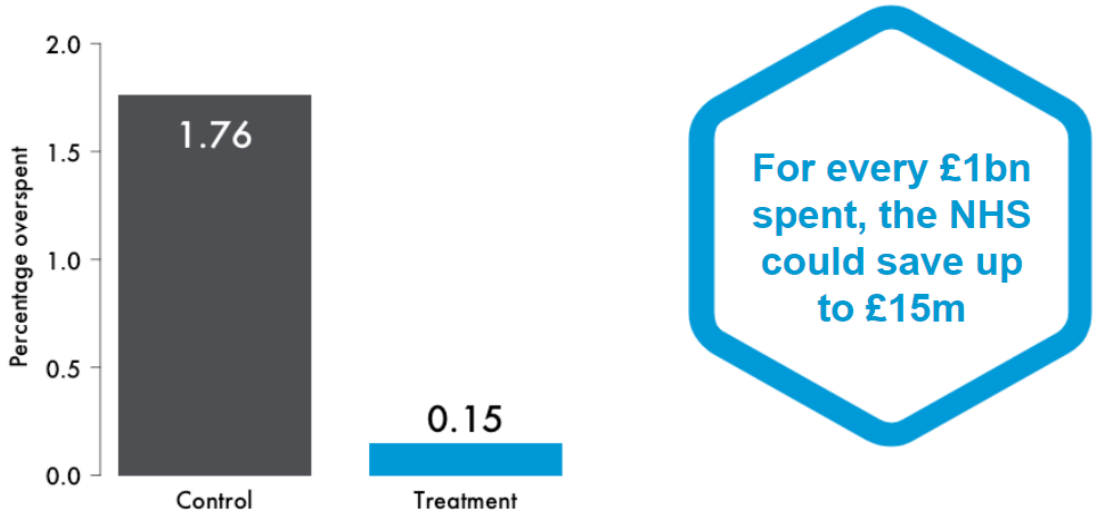
Implications

Overall, we found that our behaviourally-informed platform design helped to substantially reduce procurement overspend by reducing the amount spent on delivery and increasing the number of products bought from the cheapest supplier.

In our trial, participants in the control group spent approximately £3.5m in fictional orders across both rounds, of which £62,000 was over expenditure. In the treatment group, overspend was only £5,000 of the £3.6m spent. In percentage terms, this means that around 1.76% was overspent in the control group, compared to 0.15% in the treatment group. Therefore, our

behaviourally-informed platform design has helped to generate approximately 1.6% savings in this trial.

Figure 17. Treatment impact on savings



For NHS hospital procurement, we estimate that this would amount to savings of up to £15 million for every £1 billion spent, if the effect of the same magnitude was realised. Given Virtualstock's market share currently represents around £3.5bn of NHS procurement spend,[1] this intervention could help save around £50 million per year, should the new design be rolled-out.

However, as noted in section 5 below, there are some limitations with this experiment. For example, we have not accounted for brand loyalty, we have used an artificial shopping list, and we have not been able to account for some behavioural influencing factors such as cognitive load. It is therefore likely that our effect size may be smaller if our intervention was implemented in the real world. Despite these limitations, if the effect size in the real world was only half of that observed in this online experiment, it could still help save around £25 million per year.

Recommendations

Overall, our results show that meaningful savings can be achieved from applying behavioural insights to digital procurement platforms. This suggests that procurement platforms - in healthcare and beyond - should be designed in ways that recognise the complexity of decision-making environment and limitations of human cognition so as to enable consumers to make the most optimal purchases.

Based on the findings in this report we recommend:

[1] The NHS spends approx. £9 billion per year on procurement. Of this, around 80% - ~ £7 billion - of products and services are ordered via digital procurement platforms directly. The current market share of 'Edge for Health' by Virtualstock, the platform we targeted, is 50%, which is equivalent to £3.5 billion annually.

- ➢ **Implementing the 'order swap' element of this intervention on procurement platforms used in NHS, such as the Virtualstock platform.**
- ➢ **Evaluating this intervention in NHS Trusts.** The implementation and roll-out of this intervention should be accompanied by a robust evaluation to check whether results replicate in the field. Close cooperation with the platform owner and ongoing monitoring and updating will be key to ensure successful roll-out.
- ➢ **To continue research on hospital procurement practices, including other issues (such as errors and quality), target areas of focus (such as buyer decision-making) and solutions (such as predictive procurement)**.

1.Background

Reducing wasteful spending in hospital procurement remains one of the key challenges to improve the efficiency and financial sustainability of the NHS. According to Lord Carter's review in 2016, £700m of savings could be achieved across acute and non-acute trusts in the purchasing of general supplies and services.[2] Purchasing decisions in large NHS organisations are made by hundreds of individuals, often using complex or outdated digital systems. Small inefficiencies can compound to form large losses over time.

When it comes to the procurement of basic commodities, the inflexibility and poor user design of some existing NHS digital procurement platforms results in errors, inefficiency and overspend. Complex interfaces complicate purchasing decisions for procurers, are not designed to accommodate incremental improvements of the platform and are ill-suited to help identify the sources of spending inefficiency and variations among Trusts. Opportunities to intervene in these digital procurement systems at the point of purchase are therefore rare, but policy makers and procurement practitioners alike share the ambition for this to change.

Recent innovations, such as the arrival of Virtualstock's online procurement platform 'Edge for Health', have provided an opportunity to re-think the design of digital procurement to encourage savings by embedding behavioural insights and iterative empirical testing in the system.

In 2016, BIT received an award from the Health Foundation to explore and test ways to improve hospital procurement platforms. In the first phase of this award, BIT carried out initial research on this topic, including a desk-based literature review on procurement behaviours, qualitative research with requisitioners, led by our academic collaborators from King's College London, as well as exploration of a series of potential interventions with key agents of change in the procurement context (including Virtualstock (VS), AdviseInc. and NHS Improvement).

In the second phase of this award, BIT conducted an online experiment to test whether small changes to the digital choice architecture could alter procurement decisions. We partnered with Virtualstock (VS), the provider of an innovative new online procurement platform, to

[2] Carter Review (2016). Operational productivity and performance in English NHS acute hospitals: Unwarranted variations.

evaluate the effect of changes to their platform interface using Predictiv, BIT's in-house online experimental platform. Virtualstock are rolling out their procurement platform in over 60 NHS Trusts by 2020, and so the scope for impact is substantial. The rest of this report outlines the design and results of this trial.

2. Intervention design

We designed a new digital choice architecture for the Virtualstock hospital procurement platform, based on our diagnostic of the current platform and insights from our literature review and KCL's qualitative research. As we only had scope for a two-arm randomised controlled trial (RCT), we created a single "kitchen-sink" intervention design to compare against the current version of the platform. Our intervention version of the platform incorporated several behavioural solutions to target multiple issues.

We consulted on the design of the intervention with Virtualstock to ensure the feasibility of its potential implementation, and also with a small group of procurement officers from the NHS Sheffield Teaching Hospitals Trust. Below, we outline the key issues identified and our behavioural interventions used within the treatment.

Key issues and behavioural solutions

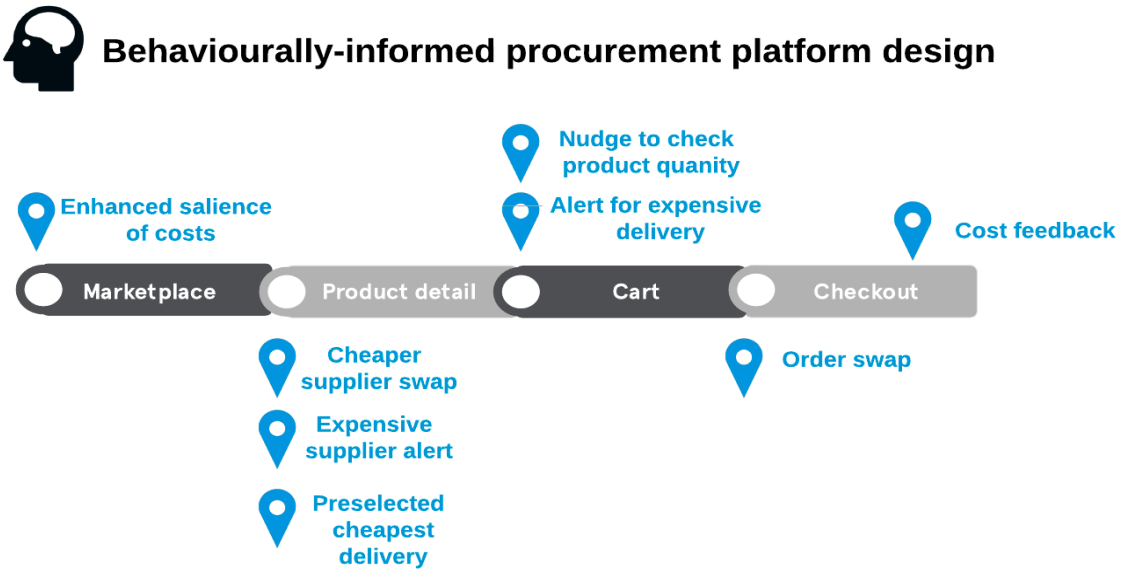
Even the most state-of-the-art procurement platforms have some design flaws that interact with finite human cognitive capacities in ways that can result in inefficiencies. We uncovered some of the main issues with the new procurement platform, drawing on design and user-centered insights from behavioural sciences, as well as observations and piloting with users. In no particular order, the key issues identified were: **complex product prices, unclear delivery options, confusion over quantities, hidden savings and lack of feedback and monitoring.**

Taken together, these issues cause complexity that may impair effective procurement. Identifying the cheapest products - taking into account the delivery, quantity and amount of information presented - is not straightforward for requisitioners and is likely to generate choice overload and decision fatigue. Whilst a lot of information is displayed, even more is out of sight; there is no information on savings that can be achieved through purchasing products in different quantities, and the defaulted delivery options are not visible. This makes finding the best deal virtually impossible, and leads to missed opportunities for savings as well as errors in items purchased and quantities. In table 2 below we summarise the key issues identified and our respective behavioural solutions.

Whilst some of the specific problems outlined here may be unique to the Edge for Health platform, many are likely to apply to other online purchasing platforms - in NHS procurement and beyond.

Figure 1 below illustrates how our behavioural solutions map onto the procurement platform process.

Figure 2. Illustration of intervention design



The solutions we developed incorporate a number of behavioural insights. In Table 2. below we summarise the key behavioural insights drawn upon in our intervention.

Table 2. Behavioural insights for procurement solutions

Behavioural insight	Detail
Defaults	Defaults are preselected options that are realised if people make no active decision. Defaults can have a significant impact on behaviour. This is in part because of people's tendency to follow the status quo and the interpretation of defaults as a signal of desirability. For instance, evidence shows the large positive impacts of 'in' defaults for a range of policies, such as organ donation or pensions uptake.[3] In the procurement context, defaults have been successfully used to encourage procurement of generic medicines. For instance, a US study found that preselecting cheaper generic medicines in the electronic prescription system increased their use by 30%, as clinicians chose to override the default only in 2% of cases.[4]

[3] Johnson, E. J., & Goldstein, D. G. (2003). Do defaults save lives?. *Science*,302, 1338-1339.
[4] Patel, M. S., Day, S. C., Halpern, S. D., Hanson, C. W., Martinez, J. R., Honeywell, S., Volpp, K. G. (2016, May 9). Generic Medication Prescription Rates After Health System – Wide Redesign of Default Options Within the Electronic Health Record. *JAMA Internal Medicine Online*. At: http://archinte.jamanetwork.com/article.aspx?articleid=2520677

Friction and active choice	Whilst simplification and defaults can help people to make a decision quickly, making an action marginally more difficult will slow them down. This can be useful when the best course of action is not self-evident and/or autonomy of decision needs to be protected. For instance, in a trial, people were more likely to sign up for vaccination when they were asked to make an active choice, out of options presented with clear cost information.[5] In the healthcare context, these insights have been used to avoid the overprescription of antibiotics. In a US study, prescriptions of antibiotics were reduced thanks to automatic prompts to provide justification or suggestions of alternative methods.[6]
Salience	The quality of decision-making in digital environments can be improved by including timely alerts - to make the best option salient, as well as to flag a potential error. For instance, in BIT's own trial, we used color-coded alerts on the e-Referrals System (eRS) to flag hospitals with short and long waiting times to help reduce GP referrals to clinics with long waiting times.[7] In the healthcare context, timely salient alerts have been used to help avoid errors and redundancy in test ordering.[8] [9]
Feedback	Making good decisions is hard when we lack a clear benchmark. Letting people know how much a behaviour costs can be such a way to help focus minds. In the healthcare context, several studies have shown that providing cost feedback flagging the price of order can significantly reduce overspend and overordering.[10] [11] [12] For instance, the orders of blood assay tests at the Nottingham University Hospital went down by a third, after the following was inserted next to the order button: "*Cost per test £1.00; total NUH spent on [item] in 2010 was £200, 914.*"[13]

[5] Keller, P. A., Harlam, B., Loewenstein, G., & Volpp, K. G. (2011). Enhanced active choice: A new method to motivate behavior change. Journal of Consumer Psychology, 21(4), 376–383.

[6] Gong, C. L., Hay, J. W., Meeker, D., & Doctor, J. N. (2016). Prescriber preferences for behavioural economics interventions to improve treatment of acute respiratory infections: a discrete choice experiment. *BMJ open*, *6*(9), e012739.

[7] The Behavioural Insights Team (2018). Update report 2016-2017.

[8] Neilson, E. G., Johnson, K. B., Rosenbloom, S. T., Dupont, W. D., Talbert, D., Giuse, D. A., ... & Miller, R. A. (2004). The impact of peer management on test-ordering behavior. *Annals of internal medicine, 141*(3), 196-204.

[9] Steele, A. W., Eisert, S., Witter, J., Lyons, P., Jones, M. A., Gabow, P., & Ortiz, E. (2005). The effect of automated alerts on provider ordering behavior in an outpatient setting. *PLoS Med, 2*(9), e255.

[10] Feldman, L. S., Shihab, H. M., Thiemann, D., Yeh, H. C., Ardolino, M., Mandell, S., & Brotman, D. J. (2013). Impact of providing fee data on laboratory test ordering: a controlled clinical trial. *JAMA internal medicine, 173*(10), 903-908.

[11] Fogarty, A. W., Sturrock, N., Premji, K., & Prinsloo, P. (2013). Hospital clinicians' responsiveness to assay cost feedback: a prospective blinded controlled intervention study. *JAMA internal medicine, 173*(17), 1654-1655.

[12] Fang, D. Z., Sran, G., Gessner, D., Loftus, P. D., Folkins, A., Christopher, J. Y., & Shieh, L. (2014). Cost and turn-around time display decreases inpatient ordering of reference laboratory tests: a time series. *BMJ quality & safety, 23*(12), 994-1000.

[13] Fogarty, A. W., Sturrock, N., Premji, K., & Prinsloo, P. (2013). Hospital clinicians' responsiveness to assay cost feedback: a prospective blinded controlled intervention study. *JAMA internal medicine, 173*(17), 1654-1655.

In the remainder of this section we provide more detail on the issues identified in turn, and the solutions designed to overcome them.

2.1 Issue 1: Complex product prices

Requisitioners can struggle to find the cheapest option when comparing identical products offered by different suppliers. We identified a number of potential issues relating to this:

- Products are ordered by cost for the total product, irrespective of the quantity it is sold in. This can skew the price perceptions: for instance, a nasal cannula sold by individual unit ranks as the first at £0.50, whilst the cheaper option per individual unit ranks second at £20 for a package of 50 cannulas. The platform does not use any common comparator - such as unit price - to help requisitioners identify the best priced options.
- Products with incomplete details, such as no photo, are ranked at the bottom, irrespective of their price.
- Delivery costs are not factored in at this point which means that requisitioners can mistakenly opt for a seemingly cheaper product that will actually be more expensive once delivery is included.

The platform does provide a box with other suppliers' offers that appears once a product has been selected. However, it includes not only cheaper but more expensive alternative suppliers.

BI Solution	Enhanced salience of cost and cheaper alternatives
	We introduced: • Price-based ordering by cost per individual unit. By default, this ordering included the cheapest delivery costs (with an option to exclude delivery, if preferred by users); • Vertical product display to prime users to focus on the top and cheapest option; • Clear display of price per individual unit; • Green highlighting and a salient tag for the cheapest supplier, defined as the supplier who sells the product for the lowest combined cost per unit of the product and the delivery charge.

Figure 3. Example of enhanced cost salience

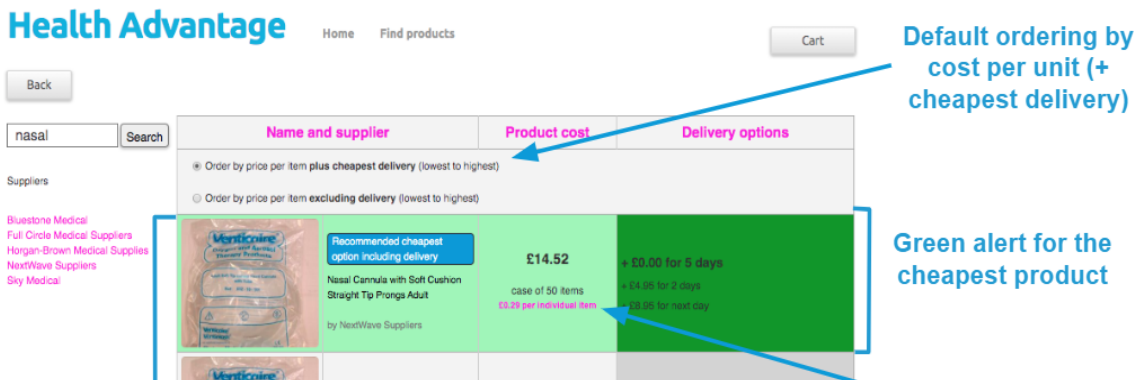
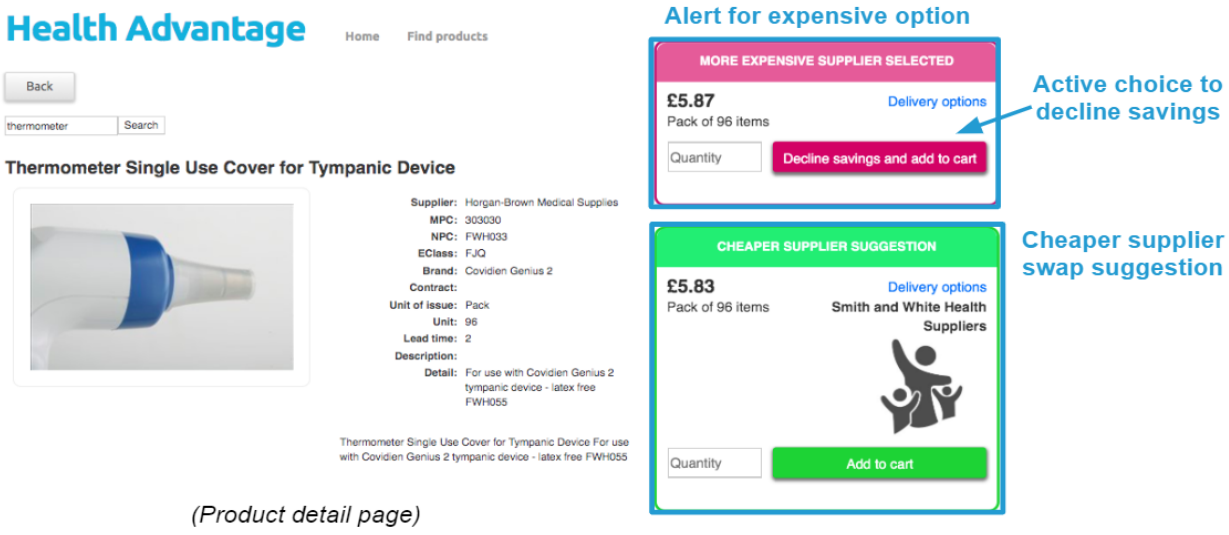


Figure 4. Example of cheaper supplier swap

2.2 Issue 2: Unclear delivery options

Suppliers choose which delivery option is the default for their products on the platform. This may be their most expensive option, such as next-day delivery. Requisitions can change the delivery option at the end during checkout, for example to a cheaper option. From a behavioural perspective, it is likely requisitioners may stick with the default delivery options, to follow the path of least resistance even though it may not be the cheapest option. To partly mitigate this issue, and based on earlier BIT's advice, the platform does include a popup alert that requires requisitioners to justify their more expensive delivery option if selected.

However, a pre-selected default delivery also means that the delivery price - and hence the total product price - is not taken into account when shopping decisions are made. This is important because the differences in delivery charges across suppliers can affect which

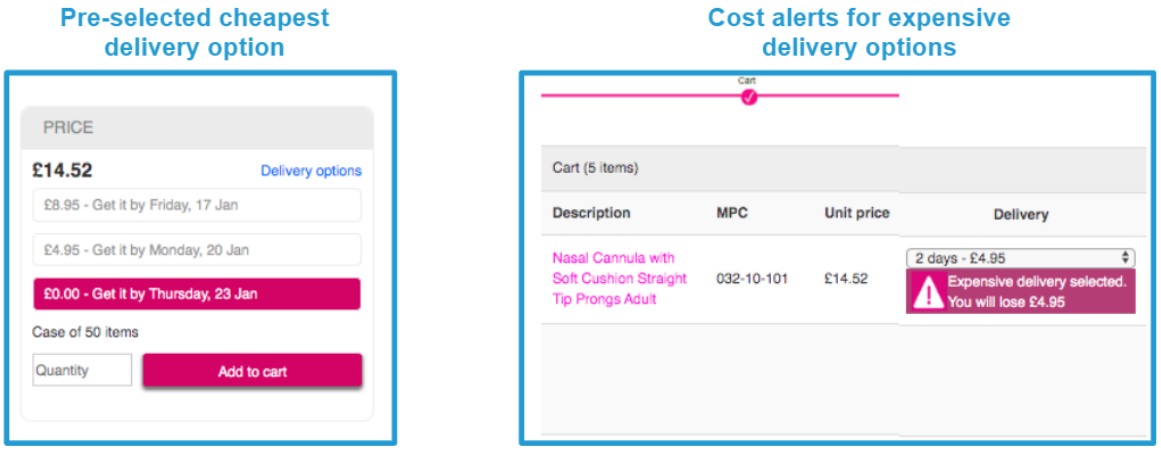


(Product detail page)

supplier is cheaper for any given product.

BI Solution	Defaults and alerts to optimise delivery
	We introduced: • Default to the cheapest delivery option for all products and suppliers; • Earlier delivery choice enabled at the point a product is added to the cart; • Salient alerts when a more expensive delivery option is selected at the cart stage and at the checkout, instead of a single pop up at the checkout.

Figure 5. Illustration of delivery default and alert



(Product detail page)

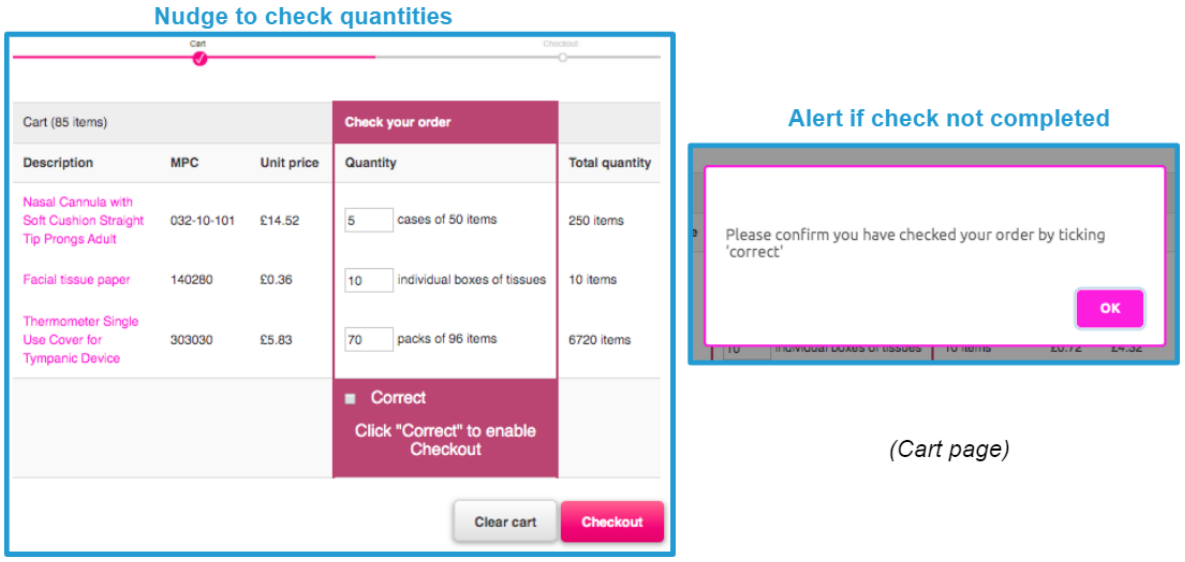
(Cart page)

2.3 Issue 3: Confusion over quantities

The size of the package of products is not clearly stated on the main search results page, in the product detail, in the cart or on the checkout page. This could lead to mistakes in order quantities but also further complicates product price comparison.

BI Solution	Nudge to check quantities
	We introduced: ● Clear information on different package sizes and the number of individual items throughout the platform; ● Primer for quantity check at the cart, encouraging users to take time, check their ordered quantities and actively confirm they have selected the correct quantity.

Figure 6. Illustration of quantity nudge



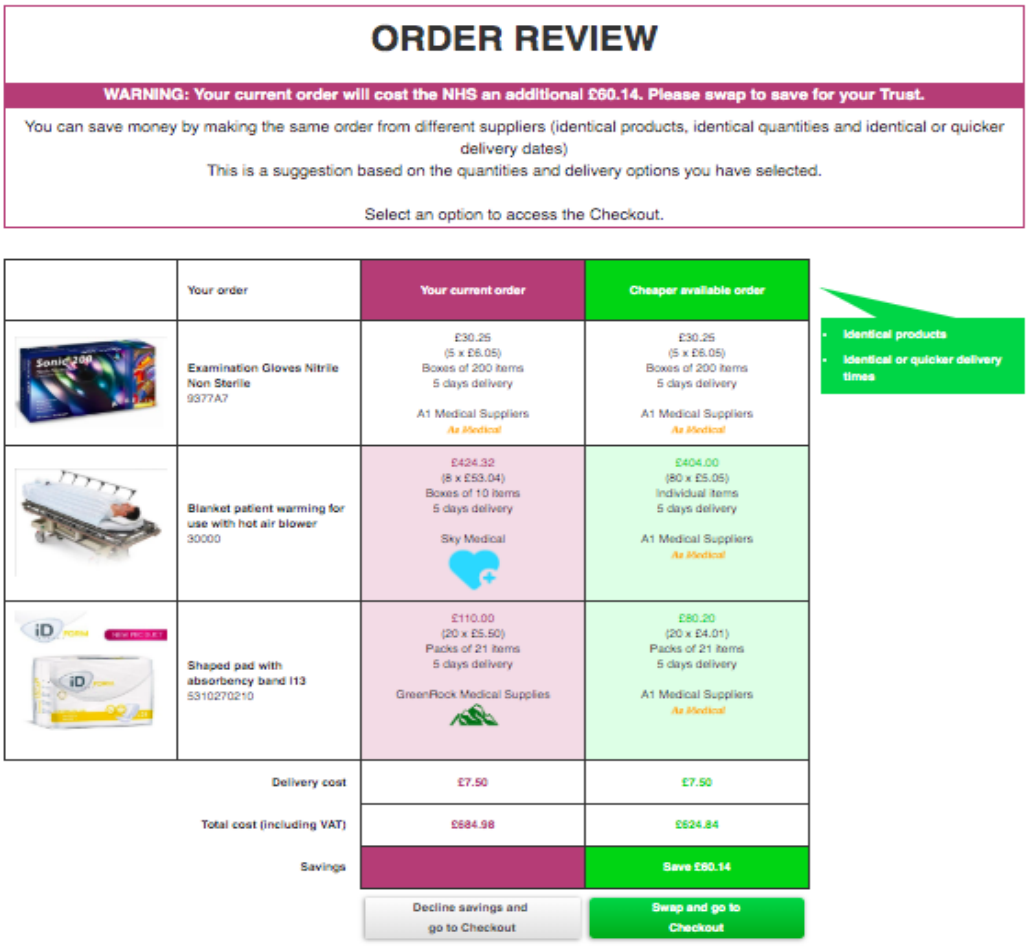
2.4 Issue 4: Hidden savings on threshold quantities and multiple items from the same supplier

Some suppliers offer special price reductions for orders depending on different quantity thresholds. However, these offers vary and are not mentioned anywhere on the portal. For instance, a supplier may offer a discount for orders above a certain quantity. In some cases it may be advantageous to order multiple items from the same provider, to save on delivery costs (provided that the product price differential does not exceed this saving).

Currently, neither of these savings opportunities are currently flagged on the platform. Requisitioners may fail to realise these savings largely due to the lack of information. Even if requisitioners were aware of such savings, the potential computational complexity of calculating different order options makes this task very difficult.

BI Solution	Order swap
	We introduced: • An ‘order swap’ suggestion with cost feedback and suggested swaps for the given basket, using loss-framing to highlight potential missed savings to the NHS. This order swap was based on an algorithm calculating the most optimally priced basket, based on the user’s selection of products, quantities and delivery dates, taking into account potential savings on large quantities and multiple item orders.

Figure 7. Illustration of order swap



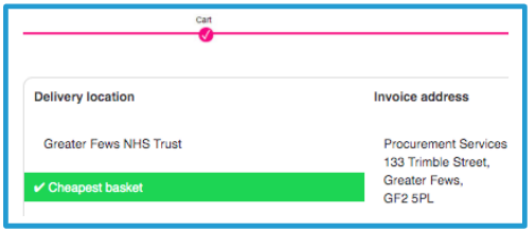
2.5 Issue 5: Lack of monitoring

On the current platform, orders have to be signed off by a supervisor. However, we know from qualitative research that supervisors often tend to approve orders in bulk, potentially without close scrutiny. Moreover, requisitioners receive little cost feedback on their spending, unless there are major concerns. This means they may remain unaware that they could have saved their NHS Trust money by buying identical products from a cheaper supplier. Similarly, they have no way of knowing how they fare compared to their peers or how they could generally improve their purchasing.

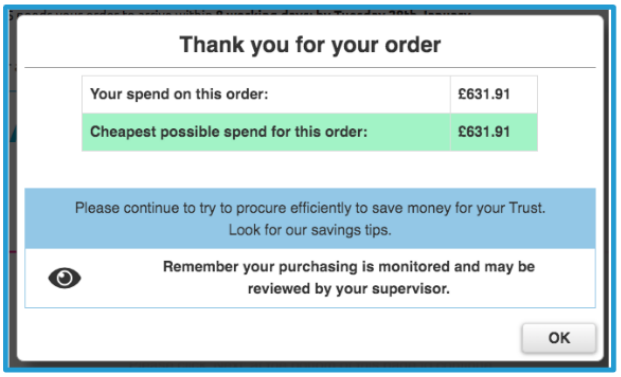
BI Solution	Cost feedback
	We introduced: ● Cost feedback on how much the requisitioner has spent on their order compared to the cheapest available basket for the identical order. To increase the sense of social accountability,

	we also included a monitoring message to highlight that the requisitioner's behaviour is being monitored and the order may be reviewed by a superior.

Figure 8. Illustration of Cost feedback



(Checkout page)



(After order page)

3. Trial design and implementation

3.1 Experiment design

Box 1. Predictiv online experiments

Predictiv (www.predictiv.co.uk) is an online research platform built by the Behavioural Insights Team to run randomised controlled trials with online populations. It enables governments and other organisations to test new policies and interventions before they are deployed in the real world.

Predictiv provides access to millions of individual participants in over 60 countries, and has the functionality to run a range of online experiments. This includes simulated choice tests, as used in this report, where individuals given a scenario and asked to make a decision. Various academic studies find that such scores correlate with actual behaviour and are more reliable than a person's self-reported understanding of the material or stated confidence.[14]

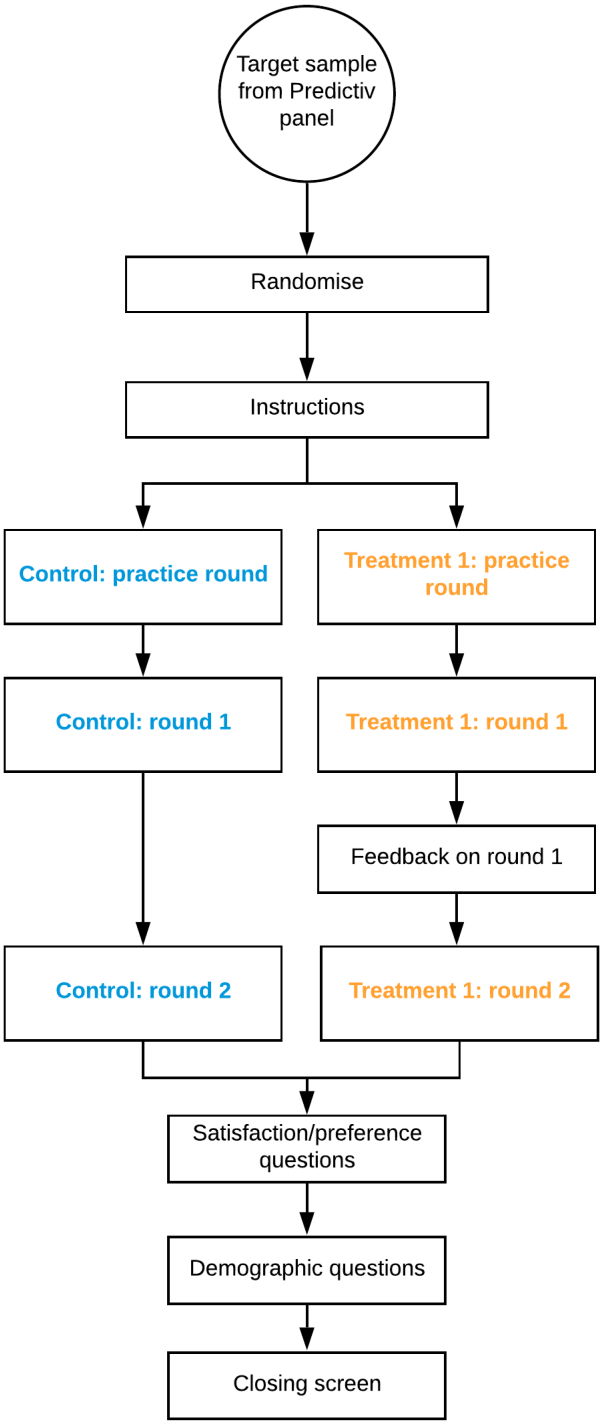
Predictiv online tests are a valuable addition to a methodological toolkit: they offer a high degree of flexibility to test different ideas; can deliver results within short timeframes; and provide quantitative, causal evidence on which ideas are most promising for changing behaviour.

Participants progressed through six stages in the online experiment run entirely on the Predictiv platform, outlined in Box 1.

First, participants answered questions to determine their eligibility for the experiment (participants needed to be aged between 18 and 99 years old and work in an NHS Trust, and a subset of participants were also required to make procurement decisions as part of their role). Participants were then asked to imagine that they worked in the procurement department for a fictional NHS Trust and had been asked to create an order using the online platform, "Health Advantage". They were also told that they were responsible for the efficient purchasing of products required by hospital staff to deliver patient care and other related services. Participants were informed that the exercise was made up of two separate shopping tasks and that there was an initial practice round.

[14] Loewenstein, G., Friedman, J. Y., McGill, B., Ahmad, S., Linck, S., Sinkula, S & Madrian, B. C. (2013). Consumers' misunderstanding of health insurance. Journal of Health Economics, 32(5), 850-862.

Figure 9. Experiment design: participant journey



Next, participants were taken through the practice round and then the two shopping rounds. After they had completed the shopping tasks, participants were asked questions on their satisfaction using the platform, whether they thought the platform could be improved, and whether they thought procurement in their own NHS Trust could be improved. Finally, participants were asked additional demographic questions including their experience of

online shopping, educational attainment, type of job within the NHS Trust, number of years working in the NHS, whether they worked in procurement, and, if they made procurement decisions as part of their role, then the number of years they made procurement decisions for.

The shopping lists in the practice round and the main rounds included instructions on what to order, the date the order was needed by , and instructions on how to use the platform. These instructions also appeared at the top of the screen when participants engaged with the platform. In the practice round, participants were asked to order 10 boxes (60 individual items) of ligation clip appliers. The shopping lists for Round 1 and Round 2 are displayed in Table 3. They were designed based on discussions with Virtualstock as well as our own pilot tests[15].

Each shopping list contains five products. We chose the products on our shopping list from a list of 15 products frequently purchased in Trusts, provided to us from Virtualstock. We chose products that varied in price and quantity. We allocated the number of items per product on the shopping list based on the typical amount of the products requested per order by NHS Trusts. This information was given to us by requisitioners as well as Virtualstock.

From discussions with Virtualstock and from piloting our experiment, we found products that are sold in multiple quantities lead to larger error rates than products sold in a single quantity. Therefore, we made sure the shopping lists each contained the same amount (three) of products that were sold in a single quantity (e.g. hearing aids y), and the same amount (two) that were sold in multiple quantities (e.g. blankets). The first three products on the shopping lists are only available in one quantity - either individual items, boxes, or packs. The final two products are sold in two different quantities - both individual items and boxes or cases.

Finally, the products on both shopping lists were required within eight working days from the day of completing the task. The longest delivery time offered on our simulation was five days. We wanted participants to feel that they could choose the longest (and often cheapest) delivery time without worrying that the product could be delayed in transit. We chose this scenario to match the real world procurement where the need for express delivery will be usually low, based on the views of the platform provider and users.

Table 3. Shopping lists for Round 1 and Round 2

Round 1
In the upcoming round, we would like you to imagine you are making a real order. Greater Fews NHS would like you to order:

[15] We ran one pilot test to assess the technical functionality of our experiment including whether the data was collecting correctly. We noticed that the shopping lists seemed to confuse participants in both the control and treatment groups as there were up to one third of participants ordering the wrong amount. We then tested different versions of the shopping list, which included changes to the format as well as changes to the amount of items required. We reduced the proportion of incorrect orders with these small changes.

- Hearing aids (MPC: 134251): **50 individual items**
- Facial tissue paper (MPC: 140280): **50 individual boxes**
- Thermometer covers (MPC: 303030): **4 packs** (96 items per pack)
- Nasal cannulas (MPC: 032-10-101): **300 individual items** or **6 cases**
- Blankets (MPC: 30000): **20 individual items** or **2 boxes**

Greater Fews NHS needs your order to arrive within **8 working days: by Wednesday 20th November.**

Round 2

In the next round, Greater Fews NHS would like you to order:

- Shaped absorbency pads (MPC: 5310270210): **4 packs** (21 items per pack)
- Blood lines (MPC: 7210697): **8 boxes** (25 items per box)
- Examination gloves (MPC: 9377A7): **4 boxes** (200 items per box)
- Oxygen facemasks (MPC: 1041): **400 individual items** or **8 cases**
- Safety cannulas (MPC: 393224): **150 individual items** or **15 boxes**

Greater Fews NHS needs your order to arrive within **8 working days: by Wednesday 20th November.**

3.2 Outcome measures

As outlined above, our goal was to design and test interventions that would improve three key areas: spend accuracy, efficiency and satisfaction. Our main focus was spend accuracy, where we primarily aimed to reduce the amount overspent on the exact same order and additionally aimed to reduce the number of quantity errors (e.g. ordering 15 items instead of 150). We were also interested in whether we could improve the efficiency of the platform and whether we could see descriptive differences in satisfaction, though these were secondary outcomes. The outcome measures for this experiment are described in Table 3.

Main outcomes

We first compared the two versions of the platform based on whether or not the participants correctly ordered products and then on whether our BI informed version of the platform could reduce the proportion and the amount of overspend for individuals who ordered correctly. We defined a correct order as an order of the five products on the shopping list in the correct quantity, without any additional or missing products. We explored these differences separately in Round 1 and Round 2, as the difficulty of the tasks were not necessarily equal and there are likely to be learning effects in Round 2.

Additional outcomes

We were also interested in the amount overspent by all participants, regardless of whether they ordered correctly; whether there was a difference in the amount overspent in Round 1 compared to Round 2; and the total amount of time spent on the platform. By assessing the amount overspent by all participants, we can extend our findings beyond our shopping lists. We can attempt to understand learning effects and the possible influence of the feedback

screen by comparing the difference in overspend between rounds. Finally, we can understand savings in efficiency by comparing the total time spent on the platform.

Table 4. Outcome measures

Main Outcomes	Target area	Definition
1. Percentage of correct orders	Spend accuracy	For each shopping round, whether the participant correctly ordered the products on the shopping list in the correct quantity.
2. Overspend on correct orders	Spend accuracy	For each shopping round, whether participants who ordered correctly overspent, and by how much they overspent by.
Additional Outcomes		
1. Overspend per item	Spend accuracy	For all participants across both rounds regardless of whether their orders were correct, the chance of overspend and the amount overspent per item ordered.
2. Difference in overspend between rounds	Spend accuracy	For all participants, the difference between their overspend in Round 2 compared to their overspend in Round 1.
3. Total time spent on the platform	Efficiency	For all participants, the total amount of time it took to complete the practice round, Round 1, and Round 2.

3.3 Sample

In total, we recruited 1,560 participants with valid responses that met our quality checks. Of those, 569 did not complete the experiment and were excluded from the analysis. The final number of participants was 991, with 491 in control and 500 in treatment. Though the amount of attrition is substantial, there was no difference in the number of participants who saw the business-as-usual platform and those who saw the treatment platform. Individuals who dropped out were similar to those who completed the experiment in terms of gender, income, and location. However, they tended to be older: 54% of those who dropped out were over 44 years old, whereas 36% of those who completed the experiment were over 44 years.

In terms of sample composition, we achieved our primary aim to obtain at least 50% of participants with procurement experience, with 64% of participants matching this criterion. Moreover, our sample is largely comparable with people who work in the NHS[16] in terms of

[16] We could not find demographic information on those who work in NHS procurement. Instead, we compare our sample with the demographics of all NHS staff.

gender, age, and income. Our sample had a larger percentage of males[17], individuals under 45 years of age[18], and a higher mean income[19]. Appendix Table A1 displays the distribution of characteristics for our sample.

The sample was balanced across the control and treatment groups for all characteristics except self reported experience in online shopping, where there were fewer participants in the control group who rarely shopped online and more participants in the control group who sometimes or often shopped online[20].

4. Results

Box 2. Summary of results

Overall, the results indicate that including behavioural insights and data-informed solutions to the Virtualstock platform can substantially improve overspending behaviour, but has little on efficiency.

The treatment substantially reduced the chance of overspend in both rounds, and reduced the average amount of overspend in Round 2. First**,** compared to the control group, the treatment group decreased the chance of overspend, both in correct orders only and in all orders made on the platform. This effect was large: the treatment reduced the chance of overspend by 91 percentage points in correct orders in Round 1, 80 percentage points in correct orders in Round 2, and 86 percentage points across all orders. Second, there was no difference between the control and treatment groups in the amount overspent in Round 1, though the amount overspent in Round 2 was £18 less in Round 2. This means that our treatment helps to prevent instances of overspend from occurring, rather than reducing the amount of overspend when it happens.

The treatment did not increase the proportion of correct orders in Round 1, and actually decreased correct orders in Round 2. Compared to the control group, the treatment decreased the proportion of correct

[17] 48% of our sample were female compared to 77% of NHS staff are female. https://www.nhsemployers.org/-/media/Employers/Documents/Plan/DIVERSITY-AND-INCLUSION/EQW19/Gender-in-the-NHS-infographic.pdf

[18] 63% of our sample were under 45 years compared with 53% of NHS staff are under 45 years. https://www.nhsemployers.org/-/media/Employers/Publications/Age-in-the-NHS.pdf

[19] The mean income for all NHS staff is £32,257 and £39,849 for professionally qualified staff. The mean income in our sample ranged £47,500 to £47,499. https://digital.nhs.uk/data-and-information/publications/statistical/nhs-staff-earnings-estimates/june-2019-provisional-statistics

[20] To control the unbalance, self-reported experience of online shopping was added as a covariate to all of our models.

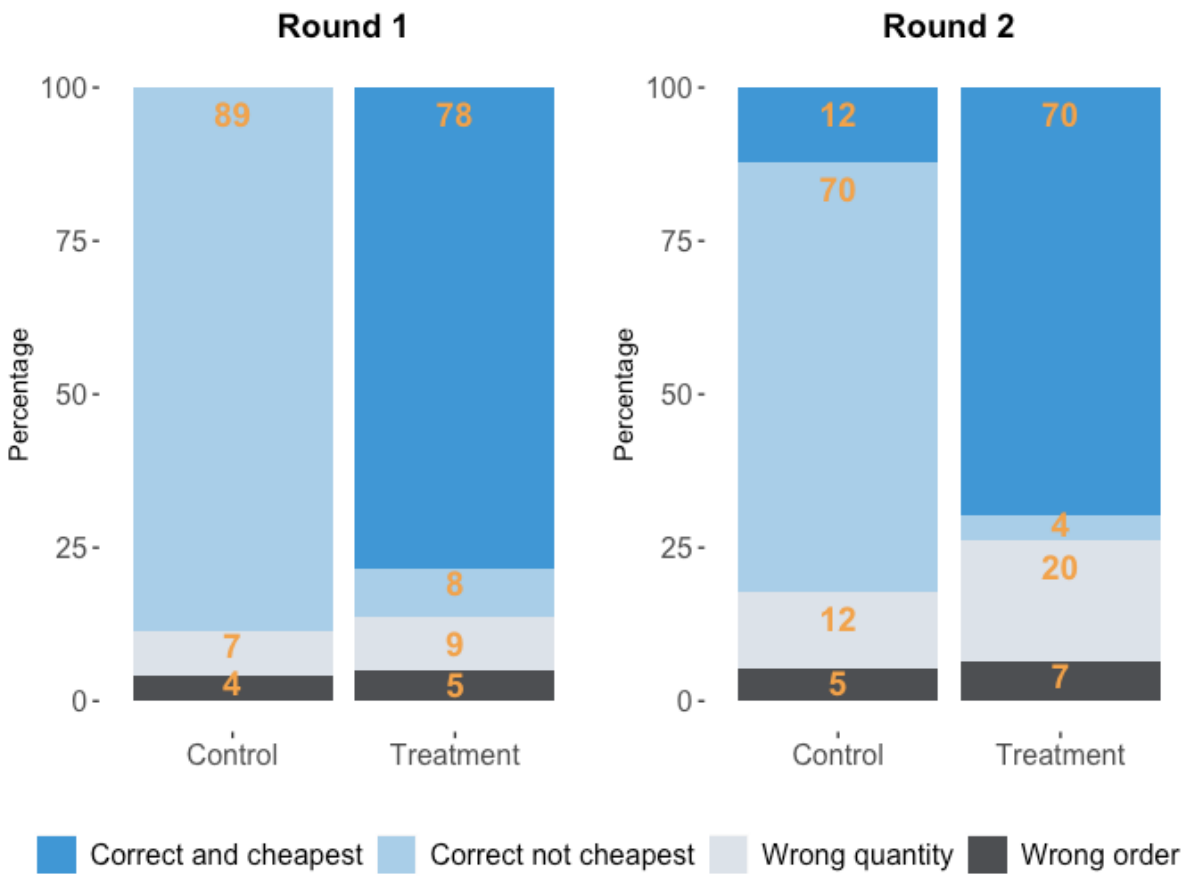
orders by 3 percentage points in Round 1 and 8 percentage points in Round 2.

There was no statistically significant difference in the efficiency of the platform. On average, participants engaged with the platform for 14 minutes in the control group and for 15 minutes in the treatment group.

This section contains graphical results of the regression analyses. Regressions are statistical tests to assess whether we can be confident that differences in participants' responses are caused by the differences in the presentation of the assessment report, rather than being due to chance. The full regression results are available in section 8.2.

The distribution of correct and cheapest orders are displayed in Figure 4.1. We see similar proportions of participants ordering correctly (i.e. ordering the exact items from the shopping list without mistakes) in Round 1 across the control and treatment groups, though fewer participants ordered correctly in the treatment group in Round 2. Many more participants ordered the cheapest basket in the treatment group in both rounds. The proportion of participants completing the wrong order (e.g. not ordering the products on the shopping list) was similar across the control and treatment groups across both rounds.

Figure 10. Distribution of answers to Round 1 and Round 2



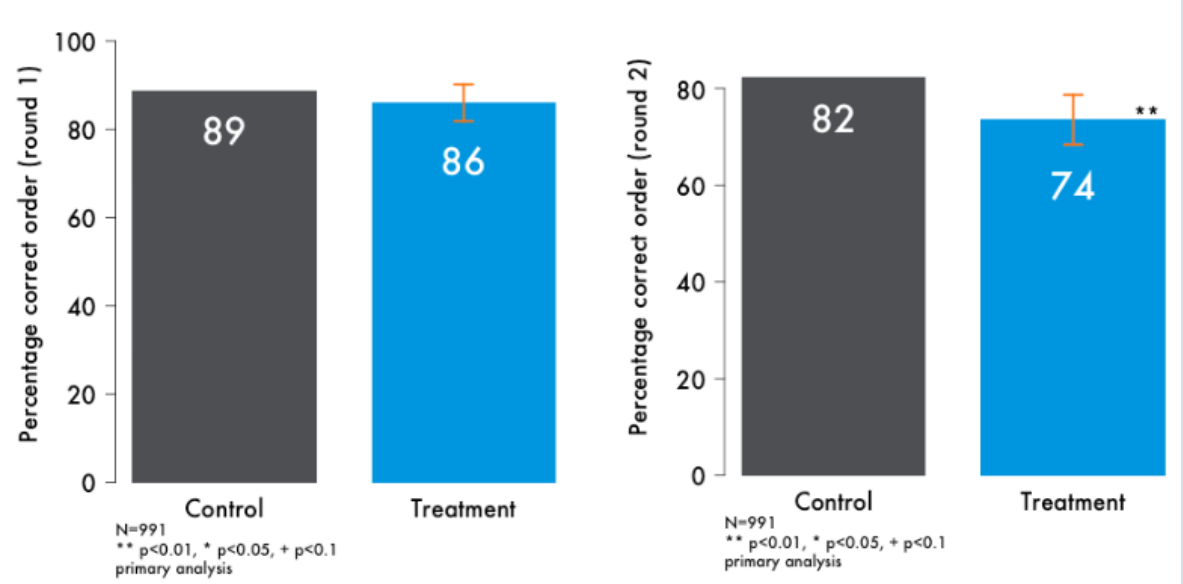
4.1 Tackling spend accuracy

We first compared the two versions of the platform based on whether or not the participants correctly ordered products and then on whether our behaviourally informed version of the platform could reduce the proportion and the amount of overspend, for individuals who ordered correctly. Though there was **no statistically significant difference in the proportion of individuals who ordered correctly in Round 1**, **fewer participants ordered correctly in the treatment group in Round 2**. We think this unexpected result could partly be an artefact of our experimental design (see Box 3. below).

Box 3. Shopping list design behind the higher number of incorrect orders in Round 2?

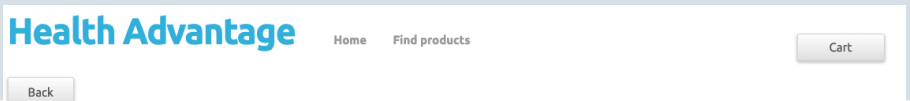
In Round 2, we observe that the majority of incorrect orders from the treatment group, compared to the control group, are quantity order mistakes for safety cannulas - the very last item on the Round 2 shopping list.

Figure 11. Percentage of correct orders in Round 1 and Round 2



76 participants in the treatment group (over 15%) ordered the wrong quantity of safety cannulas, with nearly all ordering 15 - but the shopping list requested 150. Prior to ordering safety cannulas, participants in the treatment are always presented with case-packed products as the green-highlighted search result.

Figure 12. Screenshot of Treatment, showing oxygen facemasks ordered with cases as the cheapest option.

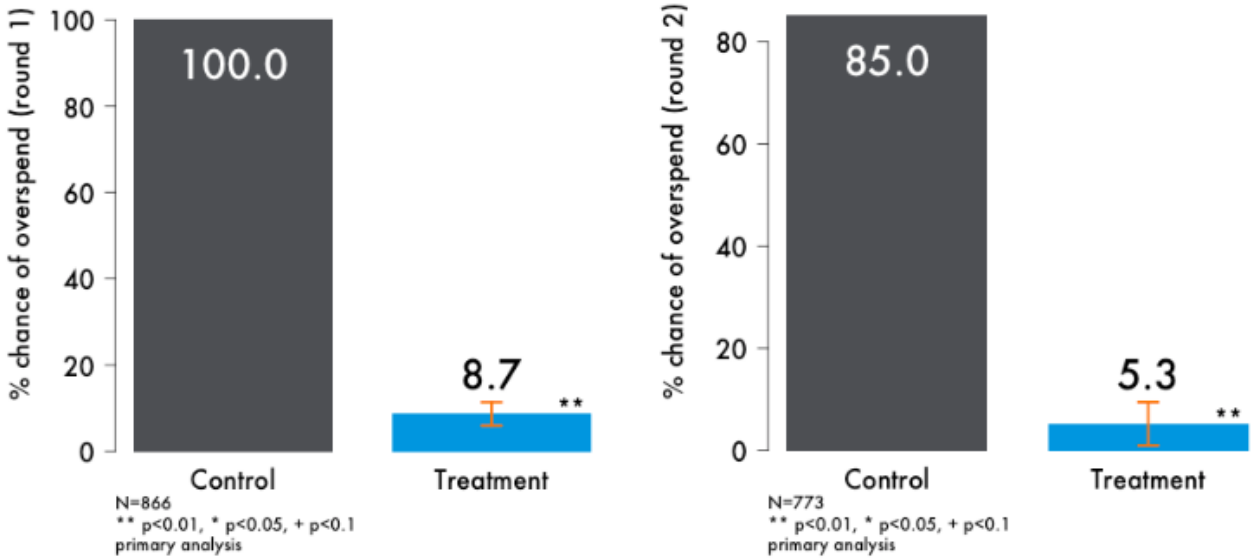


This means the relevant quantities selected for shaped absorbency pads, blood lines, examination gloves, and oxygen face masks are all single-digit: 4, 8, 4, and 8. It is perhaps unsurprising that 76 people in the treatment then mistakenly order 15 safety cannulas instead of 150, when the green-highlighted search result is listed as an individual item instead of a pack of 10 safety cannuals.

We found that the **chance of overspend was greatly reduced in the treatment group** in both Round 1 and Round 2 for individuals who ordered correctly (Figure 4.2), suggesting that our BI informed version of the platform can greatly improve spending accuracy. Compared to the control group, the chance of overspend in the treatment group was 91 percentage points lower in Round 1 and 80 percentage points lower in Round 2. Our exploratory analysis suggests this was due to the order swap element of the interventions where 96% of participants in the treatment group in Round 1 and 97% in Round 2 chose to swap to the cheapest basket for their order (see Box 3). Interestingly, we found a statistically significant association with making procurement decisions as part of their role, where those made procurement decisions as part of their role were more likely to overspend in both rounds (3 percentage points in Round 1 and 6.5 percentage points in Round 2; see Appendix 8.2.2 for full results).

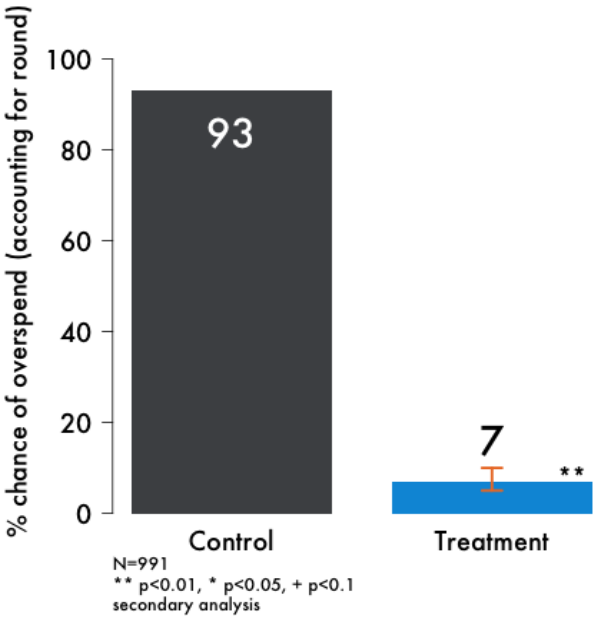
Figure 13. Percentage chance of overspend for correct orders in Round 1[21], Round 2[22]

[21] *Note*. N refers to the total number of participants who ordered correctly in Round 1 (e.g. ordered the products on the shopping list with no mistakes). Control N = 435; Treatment N = 431.
[22] *Note.* N refers to the total number of participants who ordered correctly in Round 2 (e.g. ordered the products on the shopping list with no mistakes). Control N = 404; treatment N = 369.



This finding holds when including all participants across both rounds regardless of whether their orders were correct - the chance of overspend per item ordered was 86 percentage points lower in the treatment group (Figure 4.3). Here, we found no statistically significant association between the chance of overspend and experience working in NHS procurement (full results presented in Appendix 8.2.3).

Figure 14. Chance of overspend for all orders, across both rounds



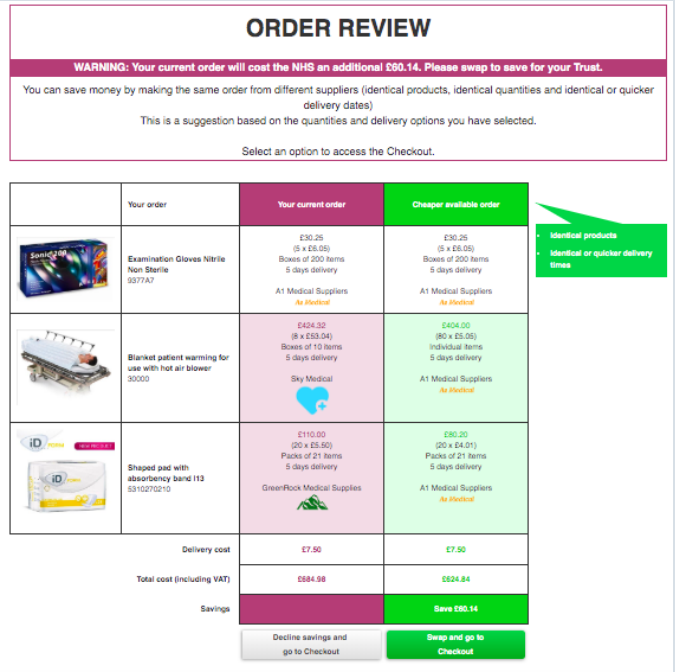
However, we cannot draw strong conclusions on the average amount overspent when it occurs, because the treatment eliminated the chance of overspend occurring so effectively. In the treatment group, only 47 individuals overspent on an order in Round 1 (39 on the correct order) and only 27 in Round 2 (21 on the correct order). **This result should be interpreted as because the effect of the treatment was so strong on *whether an***

***individual overspent*, that we cannot draw strong conclusions on the effect of the treatment on the *amount overspent, when it occurs*.** The full results from these analyses are presented in Appendix 8.2.4.

Box 4. Order swaps drive the treatment effect on overspend

To untangle the effect of our 'order review' screen from the other behaviourally-informed interventions, we compared the cost of individuals' basket when they clicked 'checkout' to the cost of the basket that they completed the order with. As the basket cost at checkout was before individuals had seen our 'order review' screen, this gives an approximation of what the cost of the basket may have been if participants in the treatment group had not been exposed to the 'order review'. This is not a perfect approximation as participants were exposed to further behavioural interventions on the final screen, after the checkout page, but it is a good estimate.

Figure 7. Illustration of order swap



We found that the mean overspend in Round 1 in the treatment group would have been approximately £224 if individuals had not been exposed to the 'order review' screen. This is in fact higher than the control group, where the mean overspend in Round 1 was £63. Similarly, in Round 2, individuals in the treatment would have overspent by approximately £47 if they had not been exposed to the 'order review' screen, compared to only £30 overspend in the control group.

This result suggests our changes to the platform before the checkout screen (e.g. the order that the suppliers are displayed on the results screen) may have actually increased overspend. This is because the earlier prompts flagged cheaper products per item, including the cheapest delivery, but did not account for the specific quantities or combinations of several goods that requisitioners may select. T**he key takeaway here is that finding the cheapest overall basket is a**

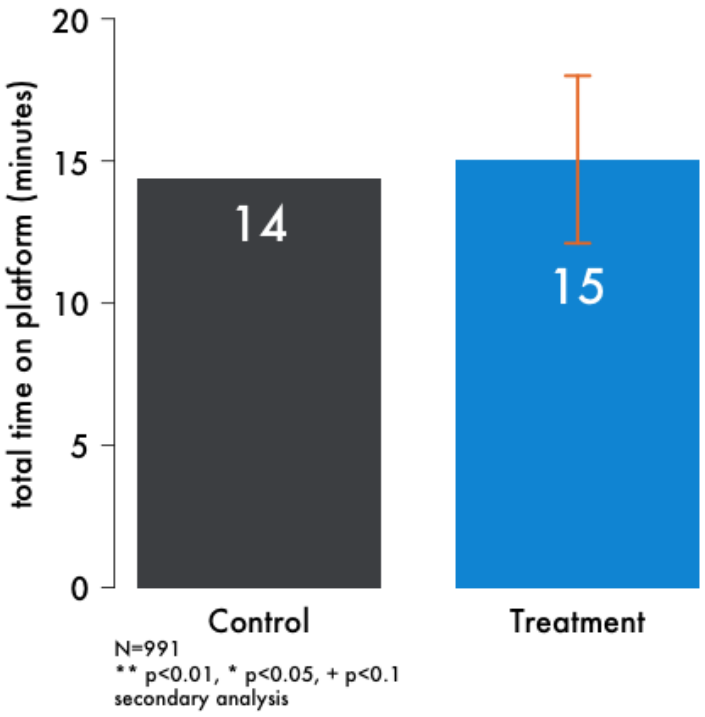
task of such a substantial computational complexity that only a sophisticated analytical prompt can help requisitioners to realise savings.

Our final assessment of whether the treatment version of the platform improved spending accuracy compared the difference in overspend between the rounds. We found that the difference in overspend was statistically significantly greater in the control (£65) than the treatment (£10) group (see Appendix 8.2.5 for full analysis). However, if we compare the percentage change in mean overspend, we see that that the treatment reduced average overspend proportionately more than the control (89% reduction in treatment vs. 67% reduction in control), suggesting that the proportion of overspend in Round 2 was reduced more in the treatment group than the control group. **Importantly, this analysis cannot distinguish whether the difference is due to differences in the learning effects between the rounds or the feedback screen shown in the treatment version of the platform.**

4.5 Tackling efficiency

On average, participants in the control group took 14 minutes to complete orders on the platform for the practice round, Round 1 and Round 2. Similarly, participants in the treatment group took on average 15 minutes. The difference between the treatment and control in the total time spent on the platform was not statistically significant (Figure 4.4), and there was no statistically significant association between experience in procurement and the time spent on the platform. **This suggests that the treatment did not improve ordering efficiency, though it did not make efficiency worse either.** This is a promising result as it shows that individuals are making better decisions while spending the same amount of time on the platform.

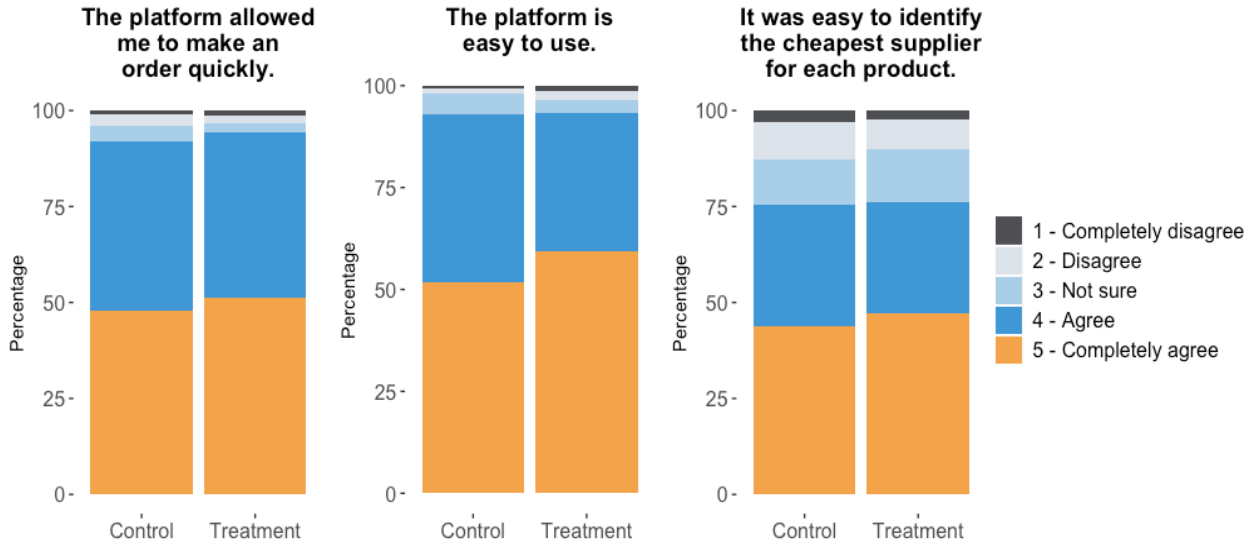
Figure 15. Total time spent on the platform



4.6 Tackling satisfaction

Following the shopping tasks on the simulated procurement platform, participants were asked for their opinions on the platform, and on the system in place at their hospital. First, participants were asked the extent to which they agreed with three statements: the platform allowed me to make an order quickly; the platform is easy to use; it was easy to identify the cheapest supplier for each product. As shown in Figure 4.5, the responses were distributed similarly across the treatment and the control groups, though more people tended to 'completely agree' with the statements in the treatment group (see Appendix 8.3 for exact figures). This suggests that the **participants were generally more satisfied with our BI version of the platform.**

Figure 16. Distribution of responses to additional platform questions



In addition, 63.1% of participants in the control group agreed that the simulated platform could be improved to make ordering easier or more efficient, compared to 55.6% of participants in the treatment group. Improvement suggestions were then provided by participants who believed the portal could be improved. Although these responses have not all been individually examined, some general findings have been made for the control and treatment groups.

In the control group, word frequencies indicated a strong suggestion for improvements related to the presentation of the delivery option, as well as information about the available delivery or the default delivery. This is understandable, since the control group's version

defaulted to delivery options chosen by the supplier, which were sometimes expensive and not salient prior to the checkout page. Meanwhile, suggested improvements made by participants in the treatment group focused more on alternative procurement considerations, other than just the cheapest, such as the quality, preferred suppliers, and reliable deliveries. Though references to quality over the price of the product were irrelevant in our simulated shopping task, it suggests that individuals are concerned about factors other than price. This could have implications for roll-out of our intervention in the real world in that requisitioners may be reluctant to make a decision based on the price for single criterion (see also the section on limitations below).

5. Limitations

We know from previous research that the findings from online simulations are generalisable to the real world (see Box 5). However, we need to also consider the limitations of our experimental design, both in terms of the limitations to running an online experiment and the limitations of our intervention design.

Box 5. External validity of online experiments

Various academic studies show that results from designs that simulate the decision environment people face in practice map on closely to behaviour outside of the experiment. This includes work on voting[23], credit card repayment decisions[24], and anti-social behaviour such as fare-dodging in public transport[25] and accepting bribes[26]. For choice tests, we build a simulated environment in which participants can make hypothetical choices, rather than relying on a person's self-reported understanding of the material, which tends to correlate poorly with actual behaviour[27].

[23] Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, *112*(8), 2395-2400

[24] Stewart, Neil. (2009) The cost of anchoring on credit-card minimum repayments. *Psychological Science*, Vol.20 (No.1). pp. 39-41

[25] Dai, Z., Galeotti, F., & Villeval, M. C. (2017). Cheating in the lab predicts fraud in the field: An experiment in public transportation. *Management Science*, *64*(3), 1081-1100.

[26] Gneezy, U., Saccardo, S., & van Veldhuizen, R. (Forthcoming). Bribery: Behavioral drivers of distorted decisions. *Journal of the European Economic Association.*

[27] Loewenstein, G., Friedman, J. Y., McGill, B., Ahmad, S., Linck, S., Sinkula, S & Madrian, B. C. (2013). Consumers' misunderstanding of health insurance. Journal of Health Economics, 32(5), 850-862.

Simulations of this kind have good generalisability in terms of direction and order. This means that we can rely on the results to tell us whether a particular treatment is more effective than the business-as-usual approach, or compared to other treatments. However, we should be cautious when generalising absolute numbers from online experiments, such as the exact proportion of people that would reduce their overspend on procurement. In most experiments simulated decisions are hypothetical, which means that, for example, not everyone who accepted the suggested cheaper basket would do this in real life (see Limitations section below). This means that we can consider the findings in the experiment indicative of the of likely direction of impact in the real world, but not the magnitude of the effect.

5.1 Limitations within the online environment

First, it is important to note that this experiment took place in a simulated online environment, where participants were not able to make actual procurement purchases. While this does not change our conclusions about whether the behaviourally-informed procurement platform design is effective, **it is possible that some factors would be different in real life which reduce the magnitude of the observed effect**. For example, we could not simulate the effect of cognitive load or external incentives. In the real world, NHS requisitioners work in busy environments with competing demands on their attention that strain their capacity to focus and perform computations to identify the best priced products. Furthermore, NHS requisitioners may have pecuniary (e.g. salary) and non-pecuniary (e.g. care for NHS, fear of detection) incentives in the real world, though incentives to procure *efficiently* are usually limited.

Second, we had to develop an **artificial shopping list and use fictional suppliers** to assess procurement behaviour. In the real world, orders are likely reorders of past products, meaning that individuals may pay less attention to the differences between suppliers and choose the suppliers that they usually order from, perhaps creating a type of brand loyalty. Qualitative evidence suggests requisitioners can feel strong brand loyalty based on years of experience with suppliers or even personal relationships they have cultivated.[28]

Finally, **our trial had high attrition**, with approximately 30% of those who started failing to complete. This experiment was a relatively long and arguably complex task, which may have contributed to the attrition rate. The direction of potential bias caused by these factors is unclear: we could have underestimated the effect of our treatment if, for example, more motivated individuals remained in the experiment or there is higher cognitive load in real life. Alternatively, we may have overestimated the effect if brand/supplier loyalty largely influences procurement behaviour.

5.2 Limitations to our intervention design

[28] Boulding, H. (2018). Requisitioner strategies in the NHS: a qualitative analysis of procurement behaviour. European Journal of Public Health, 28(suppl_4). doi: 10.1093/eurpub/cky218.112

Clearly, our interventions had a substantial effect on ordering behaviour. However, there are a number of caveats that should be considered for implementation in the real world.

1. Some elements of our intervention could backfire if implemented in isolation

As discussed above, prompts intended to flag cheaper alternatives on the search results screen and the product details screen, if implemented alone, could actually increase overspend. The reason is twofold. First, we ordered the results page by the cost per individual item plus the delivery cost. As there is usually little difference between the cost of an item across suppliers, the delivery cost is largely responsible for this order. However, requisitioners rarely order one item of a product. When ordering multiple items of a product (e.g. 150 safety cannulas; 50 hearing aids), the difference in the cost of the product across suppliers becomes much more important than the cost of delivery.

Second, our recommendations on the search results screen do not account for delivery savings when ordering multiple products from the same supplier. Whilst the first issue could be solved by prompting requisitioners to search for products with quantities, the second requires such computational complexity that only our basket swap algorithm can consistently find these savings for requisitioners.

Furthermore, the prompts and ordering of the results page could increase quantity errors. As discussed previously, we actually increased the proportion of incorrect quantity orders, which was statistically significant in Round 2. In Round 2, we found that the majority of the quantity errors were order mistakes for safety cannulas - the very last item on the Round 2 shopping list. 76 participants in the treatment group (over 15%) ordered the wrong quantity of safety cannulas, with nearly all ordering 15 - but the shopping list requested *150*. Prior to ordering safety cannulas, participants in the treatment group were always presented with case-packed products as the cheapest option (see Figure x.x in Annex). However, the cheapest supplier for safety cannulas sells them as individual items (see Figure x.x in Annex). It is possible that our ordering by cheapest supplier per item plus delivery had primed individuals in the treatment group to assume the cheapest item was always case-packed.

2. Implementation of order swaps in the real world

Our analysis suggests that the most important aspect of our intervention design is the order swap suggestion. 97% of participants in the treatment group chose to swap to the cheapest basket for their order. However, this code requires substantial computational power which takes time to run. For example, if individuals in our experiment had ordered seven products (instead of five), it would take more than three minutes to execute the code, with only 20 different products and almost 30 suppliers to choose from. In the real world, it could take much longer to execute given orders could be more than seven products and there are many more products on the platform. Effort should be given to developing an algorithm that can support NHS procurement decisions. This algorithm should be both accurate and quick to implement.

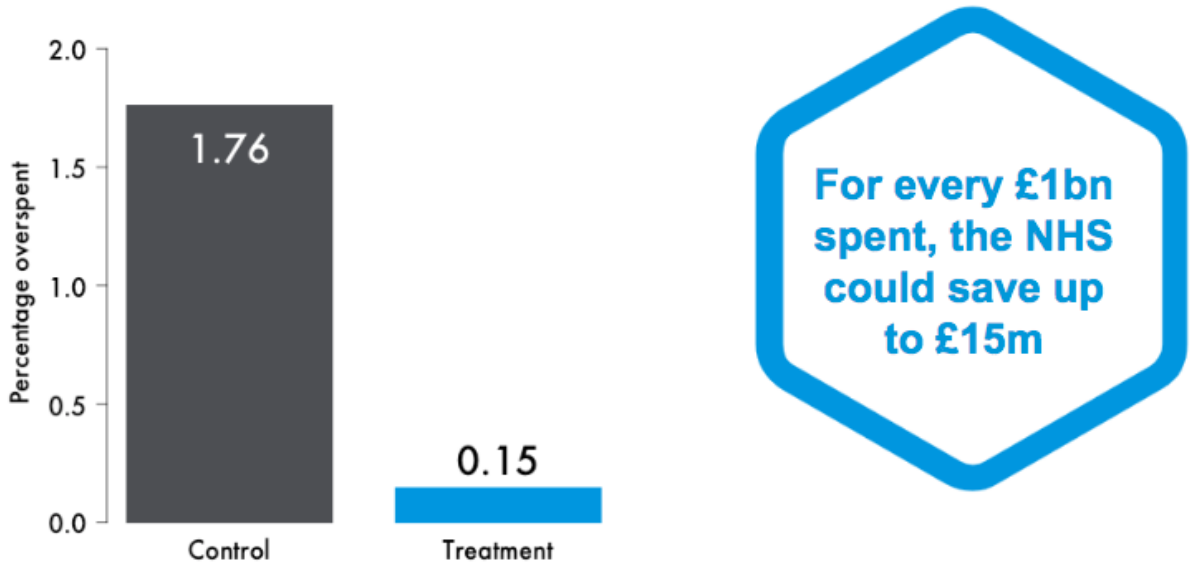
3. Focus solely on cost differences for identical products

Our treatment highlighted the costs of more expensive suppliers of identical products, using multiple alerts throughout the shopping process. Whilst price matters, supplier reliability will be as important in the real world. Requisitioners may welcome more signalling on other quality factors, for which they currently use brand as a proxy. This could come in the form of ratings from other requisitioners who have purchased a product from the suppliers.

6. Implications

Overall, we found that our behaviourally-informed platform design helps to substantially reduce the chance of procurement overspend. In our trial, participants in the control group spent approximately £3.5m across both rounds, of which £62,000 was over expenditure. In the treatment group, overspend was only £5,000 of the £3.6m spent. In percentage terms, this means that around 1.8% was overspent in the control, compared to 0.2% in the treatment group. Therefore, our behaviourally-informed platform design has helped to generate approximately 1.6% savings in this trial.

Figure 17. Treatment impact on savings



For NHS hospital procurement, we estimate that this would amount to savings of up to £15 million for every £1 billion spent, if the effect of the same magnitude was realised. Given Virtualstock's market share currently represents around £3.5bn of NHS procurement spend,[29] this intervention could help save around £50 million per year, should the new design be rolled-out.

Noting the limitations discussed above, it is likely that our effect size may be smaller if our intervention was implemented in the real world. Despite these limitations, if the effect size in the real world was only half of that observed in this online experiment, it could still help save around £25 million per year.

More broadly, these results show that such a behaviourally-informed design could be beneficial not only for NHS procurement, but might have potential to improve consumer decision-making across different public sector procurement platforms.

[29] The NHS spends approx. £9 billion per year on procurement. Of this, around 80% - £7 billion - of products and services are ordered via digital procurement platforms directly. The current market share of 'Edge for Health' by Virtualstock, the platform we targeted, is 50% equivalent to £3.5 billion annually.

7. Recommendations

In this trial, we found that a behavioural and data-informed choice architecture of digital procurement platforms can improve procurement efficiency by helping buyers to make better choices. This suggests that procurement platforms - in healthcare and beyond - should be designed in ways that recognise the complexity of the decision-making environment and the limitations of human cognition so as to enable consumers to make the most optimal purchase.

Based on the findings in this report we recommend:

- **Implementing the 'order swap' element of this intervention on procurement platforms used in NHS, such as the Virtualstock platform.** Our treatment has clear and sizable benefits in driving down procurement overspend, even when accounting for no reduction in quantity errors. Our analysis suggests this effect is largely attributable to the 'order swap' element of our intervention - the tailored algorithmic suggestion of the cheapest basket. However, the previous cost prompts could worsen overspend if implemented in isolation, as they do not take into account the overall composition of the final basket. Therefore, we recommend implementing either only the 'order swap' element or the whole intervention, but we would recommend against implementing the other prompts alone.

➢ **Evaluating this intervention in NHS Trusts.** The implementation and roll-out of this intervention should be accompanied by a robust evaluation, in order to ascertain whether the results from our online experiment replicate in the field. Such roll-out should involve close cooperation with the platform owner as well as further piloting and iteration before the implementation. Moreover, ongoing monitoring and updating of the order swap will be needed, to ensure it follows evolutions in the pricing but also other potential supplier responses to counteract it.

➢ **To continue research on hospital procurement practices, including other issues (such as errors and quality), target areas of focus (such as buyer decision-making) and solutions (such as predictive procurement)**. We recommend that future research focuses on a range of issues and explores different targets and solutions. As for issues, we need to find ways to help requisitioners avoid errors in quantities ordered but also investigate how our treatment could be adapted to be scaled to the procurement of items where other dimensions - such as quality - matter. Targeting a higher level of decision-making, such as buyers who negotiate and select available products, might bring even greater efficiency and saving gains. In any case, data-driven predictive procurement should be at the focus of any solutions as we have shown, given the computational complexity of optimising the costs. One promising area of exploration might be using data analysis of past and seasonal trends to dynamically predict the future and automate a major share of hospital procurement.

8. Appendices

8.1 Descriptive tables of sample characteristics

Table A1 reports the breakdown of the sample's demographic characteristics in the control and treatment groups, reported in percentages.

Table A1. Distribution of sample characteristics across the control and treatment groups (N = 991).

	Control (%)	Treatment (%)
Gender		
Male	52.7	51.4
Female	47.3	48.6
Age		
Under 45	62.7	63.8
45 and over	37.3	36.2
Income		
Under £32,500	20.4	21.2
£32,500 and over	75.8	75.8
Prefer not to answer	3.9	3
Location		
London	21.2	19.2
Midlands	17.3	15.2
North	24.8	25.2
South	25.9	26.8
Wales, Scotland, and Northern Ireland	10.8	13.6
Education		
Degree	83.9	80.8
No degree	16.1	19.2

Experience online shopping		
Rarely	46.2	56.4
Sometimes	36.9	30.6
Often	16.9	13
NHS procurement within role		
No	35	36.4
Yes	65	63.6
Experience working within the NHS		
Less than 5 years	17.9	20.2
5 to 9 years	19.8	16.4
10 to 14 years	19.8	23.4
15 to 19 years	19.1	19.2
20 years or more	23.4	20.8

8.2 Regression outputs

For all models outlined below, we used a step-wise approach, first including only the treatment, whether the participant had experience of procurement in the NHS, and their self-reports online shopping experience. Our second model then additionally included age, gender, income and location. Our third model included all covariates in Model 2 as well as education and number of years worked in the NHS. Here, for each research question, we report the results from the best fitting model based on the AIC statistic. All analyses use OLS models unless otherwise specified.

8.2.1 Percentage of correct orders

As outlined above, the treatment did not have a statistically significant effect on the number of participants ordering correctly (i.e. ordering the exact items from the shopping list without mistakes) in Round 1 compared to those in the control group. However, the treatment statistically significantly reduced the number of correct orders in Round 2.

As shown in Appendix Table A2, the best fitting model based on the AIC statistic was model one, including only whether someone worked in procurement for the NHS and online shopping experience. Neither of the covariates were statistically significantly associated with a chance of a correct answer. Table A3 presents model one results for the proportion of correct orders in Round 2. Again, we found that neither working in NHS procurement nor

online shopping experience were associated with a significantly different probability of ordering correctly.

Table A2. Regression results for the best fitting model on the proportion of correct orders in Round 1.

Variable	Effect
Constant	0.909 ** (0.022)
Treatment	-0.026 (0.021)
Work in NHS procurement	-0.021 (0.024)
Frequency of online shopping (compared to rarely)	
Sometimes	-0.003 (0.024)
Often	-0.047 (0.033)
R squared	0.005

Standard Errors in Parentheses, p<0.1+, p<0.05 *, p<0.01 **

Table A3. Regression results for the best fitting model on the proportion of correct orders in Round 2.

Variable	Effect
Constant	0.808 ** (0.027)
Treatment	-0.087 ** (0.026)
Work in NHS procurement	0.046 (0.029)
Frequency of online shopping (compared to rarely)	
Sometimes	-0.019 (0.03)
Often	-0.048 (0.041)
R squared	0.014

Standard Errors in Parentheses, p<0.1+, p<0.05 *, p<0.01 **

8.2.2 Percentage chance of overspend for correct orders

Table A4 below present results for the percentage chance of overspend for correct orders in Round 1. As discussed, all participants in the control who ordered the correct set and quantity of products overspent. Meanwhile, only 8.7% of such participants in the treatment

group overspent, with the vast majority (91.3%) ordering the cheapest possible basket. Interestingly, working in NHS procurement was statistically significantly associated with a 3 percentage point increase in the chance of overspending on the correct products and quantities.

Model 2 was the best fitting model for the chance of overspend in Round 2. We find that those working in NHS procurement had a statistically significantly higher chance of overspending for a correct order - by 6.5 percentage points. Furthermore, men were statistically significantly associated with a higher chance of overspending. Finally, we find that participants in South East England were also statistically significantly more likely to overspend compared to Londoners; but we do not observe any other regional differences for this outcome.

Table A4. Regression results for the best fitting model on the percentage chance of overspend for correct orders in Round 1.

Variable	Effect
Constant	1 ** (0.014)
Treatment	-0.913 ** (0.014)
Work in NHS procurement	0.03 * (0.015)
Frequency of online shopping (compared to rarely)	
Sometimes	-0.038 * (0.016)
Often	-0.034 (0.022)
R squared	0.836

Standard Errors in Parentheses, p<0.1+, p<0.05 *, p<0.01 **

Table A5. Regression results for the best fitting model on the percentage chance of overspend for correct orders in Round 2.

Variable	Effect
Constant	0.895 ** (0.038)
Treatment	-0.799 ** (0.022)
Work in NHS procurement	0.065 * (0.028)
Frequency of online shopping (compared to rarely)	
Sometimes	-0.008 (0.025)
Often	-0.013 (0.034)

Gender: Male (compared to female)	-0.046 * (0.023)
Age: 45 and over (compared to under 45)	-0.029 (0.023)
Income (compared to under £32,500)	
£32,500 and over	-0.01 (0.029)
Prefer not to answer	-0.278 ** (0.064)
Location (compared to London)	
Midlands	-0.017 (0.035)
North England	0.004 (0.033)
South England	-0.09 ** (0.032)
Wales, Scotland, and Northern Ireland	-0.014 (0.04)
R squared	0.652

Standard Errors in Parentheses, p<0.1+, p<0.05 *, p<0.01 **

8.2.3 Percentage chance of overspend for all orders

Here, we used a multilevel OLS model to assess the overspend across all orders. As we expect from the Round 1 and Round 2 results, the treatment has a statistically significant effect on the overspend per item when including all participants (regardless of whether they ordered the shopping list items). We found no significant association in NHS procurement experience.

Table A6. Regression results for the best fitting model on the percentage chance of overspend for all orders across both rounds.

Variable	Effect
Fixed effects	
Constant	0.97 ** (0.01)
Treatment	-0.86 ** (0.01)
Round 2	-0.09 ** (0.01)
Work in NHS procurement	0.02 (0.01)
Frequency of online shopping (compared to rarely)	
Sometimes	-0.02 (0.01)

Often	-0.01 (0.02)
Random Effects	**Standard deviation**
Participant level	0.02 (0.13)
Residual	0.05 (0.21)

Standard Errors in Parentheses, p<0.1+, p<0.05 *, p<0.01 **

8.2.4 Assessing the amount overspent

Gamma models were used to assess the amount overspent in Round 1 and Round 2. The sample N in this context is the number of participants who made a correct order, but overspent on it. As such, given the selection effects we have just seen the N is a lot smaller in the treatment group (N = 39) than the control group (N = 435). For those that did overspend, the amount of overspend was similar between treatment and control in round one, but in round two the treated individuals tended to overspend less.

Model 3 (including all covariates) was the best fitting model for Round 1, and the results are presented in Figure 8.2 and Table A7. We find that working in NHS procurement, online shopping experience, age, gender and region were not associated with significant differences in the amount overspent. However, we find that participants with higher income (£32,500 and over) overspent around £6.00 less than those earning below £32,500, and this difference was statistically significant.

In round 2, the best fitting model based on AIC was model 1. Figure 8.2 and Table A8 shows that the treatment significantly reduced the amount overspent for those that overspent. Specifically, we find that those in the treatment who made a correct order in round 2, but overspent, spent around £5.00 less than the equivalent group in the control. Neither working in NHS employment nor online shopping experience were strongly associated with a difference in the amount overspent.

Figure 18. The amount overspent (£) for correct orders in Round 1[30] and Round 2[31]

[30] *Note.* N refers to the number of participants who ordered correctly and overspent (e.g. ordered the shopping list but did not order the cheapest basket). Control N = 435; treatment N = 39. The large confidence interval is due to the small proportion of participants in the treatment group that did not order the cheapest basket.

[31] *Note.* N refers to the number of participants who ordered correctly and overspent(e.g. ordered the shopping list but did not order the cheapest basket). Control N = 344; treatment N = 21. The large confidence interval is due to the small proportion of participants in the treatment group that did not order the cheapest basket.

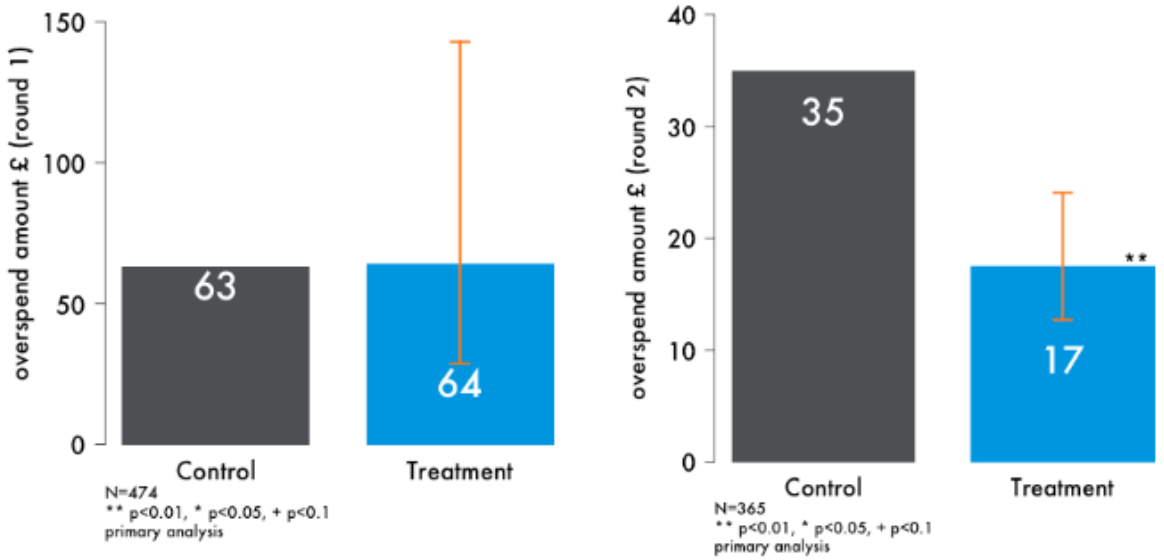


Table A7. Regression results for the best fitting model on the amount of overspend for correct orders in Round 1.

Variable	Effect
Constant	4.616 ** (0.431)
Treatment	0.016 (0.41)
Work in NHS procurement	-0.01 (0.287)
Frequency of online shopping (compared to rarely)	
Sometimes	-0.078 (0.256)
Often	0.169 (0.343)
Gender: Male (compared to female)	0.126 (0.243)
Age: 45 and over (compared to under 45)	-0.021 (0.283)
Income (compared to under £32,500)	
£32,500 and over	-0.947 ** (0.314)
Prefer not to answer	-2.505 ** (0.662)
Location (compared to London)	
Midlands	0.309 (0.368)

North England	-0.362 (0.341)
South England	0.592 + (0.34)
Wales, Scotland, and Northern Ireland	-0.429 (0.425)
Education: No degree (compared to having a degree)	-0.542 + (0.321)
Experience in NHS (compared to less than 5 years experience)	
5 to 9 years	0.252 (0.362)
10 to 14 years	0.131 (0.376)
15 to 19 years	0.331 (0.401)
20 years or more	-0.003 (0.424)
R squared	0.135

Standard Errors in Parentheses, p<0.1+, p<0.05 *, p<0.01 **

Note. In a Gamma regression the parameters are interpreted as log-multipliers. So a parameter of -2.5 means multiply the outcome by exp(-2.5). If beta is >0 it means the effect is to increase the outcome, if beta<0 the effect is to reduce the outcome.

Table A8. Regression results for the best fitting model on the amount of overspend for correct orders in Round 2.

Variable	Effect
Constant	3.496 ** (0.069)
Treatment	-0.692 ** (0.163)
Work in NHS procurement	0.15 + (0.084)
Frequency of online shopping (compared to rarely)	
Sometimes	-0.085 (0.088)
Often	-0.069 (0.113)
R squared	0.067

Standard Errors in Parentheses, p<0.1+, p<0.05 *, p<0.01 **

Though we tried to model the effect of the treatment on the amount overspent across all orders, our multilevel gamma model failed to converge. We believe this is due to the small number of participants in the treatment group that overspent at all. We managed to run a model for Round 1 only (N = 538; control N = 491; Treatment N = 47) and found no

statistically significant effect of the treatment (Table A9). However, only 27 participants in the treatment group overspent in Round 2, and our model could not run due to the small sample. As discussed above, this result should be interpreted as because the effect of the treatment was so strong on *whether an individual overspent*, we cannot draw strong conclusions on the effect of the treatment on the *amount overspent, when it occurs.*

Table A9. Regression results for the best fitting model on the amount of overspend across all orders in Round 1.

Variable	Effect
Constant	-2.058 ** (0.457)
Treatment	0.132 (0.415)
Work in NHS procurement	0.504 + (0.299)
Frequency of online shopping (compared to rarely)	
Sometimes	-0.32 (0.268)
Often	1.234 ** (0.349)
Gender: Male (compared to female)	0.17 (0.252)
Age: 45 and over (compared to under 45)	-0.184 (0.295)
Income (compared to under £32,500)	
£32,500 and over	-0.984 ** (0.323)
Prefer not to answer	-2.562 ** (0.694)
Location (compared to London)	
Midlands	0.37 (0.381)
North England	-0.158 (0.354)
South England	0.678 + (0.356)
Wales, Scotland, and Northern Ireland	-0.334 (0.445)
Education: No degree (compared to having a degree)	-0.548 (0.339)
Experience in NHS (compared to less than 5 years experience)	
5 to 9 years	0.463 (0.386)
10 to 14 years	-0.15 (0.397)

15 to 19 years	0.206 (0.419)
20 years or more	-0.165 (0.446)
R squared	0.194

Standard Errors in Parentheses, p<0.1+, p<0.05 *, p<0.01 **

8.2.5 Assessing the difference in the amount overspent between rounds

This analysis compares the control and treatment on the absolute monetary difference in the amount overspent between the rounds. These results show that in the control group, participants overspent by an average of £65. In the treatment group, this was only an average of £10, which was statistically significantly less (Table A10). However, the percentage change in mean overspend between rounds 1 to 2 was proportionately lower in the treatment than the control (89% reduction in treatment vs. 67% reduction in control). This metric could provide more reasonable support to the hypothesis that learning/feedback effects were greater in the treatment group.

Table A10. Regression results for the best fitting model on the amount overspent across the rounds.

Variable	Effect
Constant	81.982 * (32.882)
Treatment	-54.962 ** (18.572)
Work in NHS procurement	21.525 (23.321)
Frequency of online shopping (compared to rarely)	
Sometimes	11.501 (21.498)
Often	21.6 (28.739)
Gender: Male (compared to female)	31.986 (20.032)
Age: 45 and over (compared to under 45)	-10.98 (19.566)
Income (compared to under £32,500)	
£32,500 and over	-77.862 ** (24.503)
Prefer not to answer	-108.591 + (56.284)
Location (compared to London)	

Midlands	15.162 (31.247)
North England	-12.393 (28.488)
South England	54.477 + (28.164)
Wales, Scotland, and Northern Ireland	-17.821 (34.003)
R squared	0.031

Standard Errors in Parentheses, p<0.1+, p<0.05 *, p<0.01 **

Endnotes