Label-checking strategies to adapt behaviour to design

James H. Smith-Spark
smithspj@lsbu.ac.uk

Hillary Katz
katzh@lsbu.ac.uk

Alexander Marchant
marchaa4@lsbu.ac.uk

Thomas Wilcockson
wilcockt@lsbu.ac.uk

Dept. of Psychology, London South Bank University, 103 Borough Road, London, SE1 0AA, UK.

ABSTRACT

Despite robust quality control procedures, labelling errors on fresh produce are estimated to cost the UK supermarket industry approximately £50million pounds per year in product recalls and wastage. Changing the format of the labels themselves is not a viable option. Instead, the challenge is to change or guide human operatives' behaviour so that label printing errors do not go undetected during quality control procedures. To this end, a simulated label checking task was presented to naïve participants to compare more systematic and strategic methods of label checking. Two conditions in which behaviour was computer-led were compared with a control condition in which checkers adopted their own idiosyncratic checking method. The data indicate that the two computer-led approaches resulted in improved levels of accuracy. Pushing label checkers towards a more systematic approach would appear to be effective in reducing undetected label errors, and could lead potentially to significant financial savings and reduced environmental wastage in the fresh produce industry.

Author Keywords

Cognitive load; Distributed cognition; Quality control checking

ACM Classification Keywords

Ergonomics; Psychology; Process Control; Automation

INTRODUCTION

Errors in the labelling of fresh fruit and vegetables are estimated to cost the supermarket industry in the UK approximately £50million per annum (S. Hinks, Product Technical Manager: Fruit and Floral, Sainsbury's Supermarkets Ltd, personal communication). As well as the financial costs incurred in recalling wrongly labelled fresh produce from distribution depots or removing it from retail shelves, there is also a considerable cost to the environment. Perfectly good, yet incorrectly labelled, fruit

and vegetables are wasted, to say nothing of the carbon footprint which is left by having to recall and replace the produce. These labelling errors occur despite advanced IT systems and multiple quality control checks by experienced human operatives. The challenge is to understand why quality control checkers fail to notice label errors which are very obvious and, on the basis of this understanding, consider how such errors can be minimized. The research described in this paper was undertaken to compare the effectiveness of two different methods of label-checking in improving the accuracy of performance. The two methods were designed to guide label checkers towards a more systematic approach to label-checking in which they checked one field of information at a time in an order specified, to a varying degree, by a computer. Before the theoretical underpinnings of this approach are considered, it is first necessary to contextualize the research by describing the applied setting in which these label-checking errors arise.

Orders for fresh produce are communicated by the supermarket to account managers in the packaging facility's commercial office. They record which fruit and vegetable lines are required and any promotional offers that the supermarket wants to put in place for its products. This information is then compiled into a specification sheet by the commercial office team. Each week, a new specification sheet is published and distributed to workers in the pack-house, with updates occurring throughout the week. This specification sheet is the sole source of information from which pack-house operatives work to prepare and label orders as they are received. Due to unpredictable variation in some of the features of the product (such as its grower, size, and promotional offers), the full and detailed information required for the labels is not available until the produce actually arrives onsite. This means that labels must be printed and checked minutes before the packaging process begins.

The quality control procedure requires one or more dedicated operatives to check that the information printed on the product label (and any associated promotional ribbon or sticker) matches the entry for the product on the specification sheet. Depending on the type of produce, the number of fields of information which need to be checked on a label varies from three to eleven. Up to four independent checks of the label are carried out between the point at which it is packaged and it being loaded for shipment to distribution depots around the UK. Typically,

checks occur at the point the label is printed (being checked independently by both an operative in the print room and the team leader of the line responsible for packing the run), again after the first few labels have been attached to the product, and once more at the end of the packaging run.

Despite these rigorous quality control procedures, labelling errors still escape notice and lead to costly recalls of products. The frequency with which such undetected errors occur is low (affecting around 1% of label runs) but the financial and environmental consequences attached to even this small number are unacceptable and drive research to reduce their number even further. The complexity and diversity of the internal and external systems involved in producing the labels mean that it is not feasible to change the label printing process itself. Instead, it is the behaviour of label-checkers which needs to be understood and, if suboptimal approaches to checking are identified, for the behavior responsible to be adapted in such a way as to increase its effectiveness.

A preliminary investigation was undertaken to explore how individuals went about checking labels. A simulated label-checking task was administered to both experienced pack-house quality control operatives and novice undergraduate students who were naïve to the quality control checking of fresh produce labels. This paradigm was developed to be as close to verisimilitude as was possible under laboratory conditions. Using eye-tracking technology, this study showed that individuals used different (and often idiosyncratic) styles of checking to determine whether or not labels were correct. These individual differences were found in both experienced operatives and undergraduate students. Some participants were found to adopt a systematic approach to checking, tending to read one bit of information at a time from the specification sheet and checking this against the entry on the label. Other participants used a chunking approach [1], in which they read and committed to memory several bits of information at a time before checking each of these bits against the label in a single visual pass. A final group of participants seemed entirely haphazard in their approach to checking, with no obvious strategy or system being manifested in the pattern of their eye movements. Those participants who used a systematic approach to label-checking were found to be significantly more accurate in detecting label errors than those who adopted either a chunking or haphazard approach to the task.

Onsite observations of the checking process and interviews with key operatives suggested that a number of cognitive demands made on label-checkers, such as attention-switching and working memory, might also be connected with how well they performed and provide a basis for positive behavioural interventions. Working memory [e.g., 2, 3] is a system involved in the temporary storage and processing of information. It is divided into at least two modality-specific slave systems. The phonological loop is responsible for dealing with phonologically-based information (such as speech or text), whilst the visuospatial sketchpad deals with visual and spatial information (such as patterns or sequences). These slave systems are overseen by a central executive, responsible for the controlled and strategic allocation of attentional resources to the task at hand. By distributing the task over both internal and external representations [4, 5], working memory load could be reduced by offloading it to the surrounding environment [6, 7].

The approach taken in the present research to improve the detection of label errors, then, was to develop a distributed cognitive system [8] for label-checking which coupled [9] the human label checker with a computer to varying extents. Performance using these methods was compared with a control condition in which participants were allowed to develop and follow their own method of label-checking.

The first computer-coupled method (the "Systematic" approach) guided the participant's attention by highlighting information on the specification sheet with a red box, with one field of information being highlighted at a time. It is argued that by encouraging the checker to adopt a systematic, serial approach to checking, this method would direct and hold attention and thereby enhance the detection of errors.

The second approach (the "Distributed approach") involved participants receiving an auditory field-by-field presentation of the information contained on the relevant line of the specification sheet. The fields of information were "read out" to the participant in a fixed order and the participant was not permitted to progress to the next field until he or she had verified the current field was correct. This method meant that the participant could be entirely focused on the label itself and alleviated the need to refer back-and-forth visually to the specification sheet. In this way, the cognitive process of maintaining in working memory the information taken from the specification sheet would be replaced by an auditory perceptual process, akin to the epistemic actions identified in Tetris players who changed mental rotations into visual rotations [10], thereby reducing cognitive load [11]. This would be expected to reduce cognitive load and improve the accuracy of performance, by reducing the cognitive effort needed to encode and maintain a short-term memory representation of the information set out in the specification sheet. Like the first approach, this method also ensured that a systematic, serial checking method was adopted since the participants were unable to continue to the next field until they had heard all of the information for the current field.

For the reasons outlined above, it was hypothesized that the imposition of a systematic checking strategy on the participants, through methods which highlighted information visually or verbally, would improve error detection compared with a control group whose approach to checking labels was unguided and not assisted in any way.

METHOD

Participants
Fifty-three university students (37 female, 5 male; mean age = 23 years, SD = 6) took part in the experiment. All of the participants reported themselves to be naïve to the label checking process. The participants were either native English language speakers or else were studying at undergraduate level with an International English Language Testing System (IELTS) score of at least 6.0 (the minimum standard required for entry to London South Bank University's undergraduate degree programmes). They were assigned randomly to one of three experimental conditions. The background characteristics of the participants in each condition who met the criteria for inclusion in the analyses are displayed in Table 1. The conditions did not differ significantly in age, $F(2, 43) < 1$, $MSE = 38.594$, $p = .504$, or gender, $\chi^2(2) = 2.22$, $p = .329$. The participants were awarded course credits or a small honorarium for their participation.

Materials
A head-rest was employed to minimize participants' head movements during the experiment. Experimenter Builder Version 1.4.128 B (SR Research Ltd., Ontario, Canada) was used to program and run the experiment.

Facsimiles of the product specification sheet and labels were produced. The number of fields of information per product on the specification sheets and labels was held constant at seven. These fields were the product (the type of fruit or vegetable, e.g., baby courgettes), its country of origin, the grower (the name of the company which grew and shipped the product), the quantity of items contained in the packet (i.e., the weight of the product), its best-before date (denoted by "BB" on the specification sheet), the product's barcode number, and details of any promotion ribbon/label to be appended to the packaging (i.e., any promotional activity being offered by the supermarket on the product, such as "Any 2 for £2.50"). Fifty different labels were presented on an 18" VDU in each of the three blocks of trials. Figures 1 and 2 show an example of a specification sheet and a product label respectively.

Design
The two sources of information were shown at the same time, with the specification sheet appearing in the top half and the label in the lower half of the computerized display.

Condition	N	Mean age (years)	Gender
Control	12	26 (SD = 7)	9F, 3M
Systematic-first	18	23 (SD = 5)	16F, 2M
Distributed-first	16	23 (SD= 7)	15F, 1M

Table 1. Background characteristics for each condition.

Figure 1: An example of a specification sheet.

In each block of 50 trials, there were 40 trials where the information presented on the specification sheet and the label matched and 10 trials where it did not. On any given mismatch trial, only one field of information out of the seven differed between the two sources. This was varied semi-randomly over trials so that errors occurred in different fields. Correct responses to these trials occurred when a mismatch between the information on the label and that on the specification sheet was detected and reported.

A mixed-measures design was used, with the between-subjects factor being checking condition (levels of treatment: control, systematic-first, and distributed-first) and the within-subjects factor being experimental block (levels of treatment: Block 1, 2, and 3). The dependent variable was the proportion of correct responses made by the participants and was analyzed using 3 x 3 mixed-measures analysis of variance (ANOVA).

Procedure
The participants gave informed consent to take part in the experiment. Prior to the start of the checking task, a 10-minute slide show presentation was given to all of the participants, giving a detailed description of the label layout, specification sheet layout, general task instructions, the nature of errors, etcetera.

In Block 1, all of the participants adopted their own idiosyncratic approach to label checking, having received no specific instructions about possible strategies that might

Figure 2. An example of a product label.

be used. The participants were asked to make a decision as to whether or not a given label was correct by pressing designated Yes and No keys on a standard QWERTY keyboard.

The participants allocated to the control condition continued to use their own self-chosen method of checking through Blocks 2 and 3.

By contrast, one of the two experimental groups was shown and instructed to use a 'Systematic' method of checking prior to the onset of Block 2. This entailed moving a red box which highlighted the current field of information across the specification sheet as they checked it against the produce label.

Before the start of Block 2, the other experimental group was shown and instructed to use a 'Distributed' method of checking which was similar to the Systematic approach in having the fields of information on the specification sheet highlighted sequentially by a red box but, in addition, a pre-recorded voice read out the same bit of information.

In Block 3, the experimental groups switched methods, so that participants who had employed the Systematic approach in Block 2, next were shown and instructed to use the Distributed approach in Block 3, whilst participants who had employed the Distributed approach in Block 2 where shown and instructed to use the Systematic approach in Block 3.).

The participants were given a written debrief at the end of the experiment.

RESULTS

Seven participants were excluded due to proportion accuracy levels of < .60 to mismatch trials in Block 1, indicating that they were performing at or near chance on indicating mismatches. The details reported in the Participants subsection reflect the reduced data set once these participants had been removed.

Overall, the three conditions performed at a very similar level of accuracy (Control mean = .88, *SEM* = .01; Systematic mean = .88, *SEM* = .01; Distributed mean = .88, *SEM* = .01). The two-way ANOVA indicated that there was no significant effect of condition on overall accuracy of responses, $F_{(2, 43)} < 1$, *MSE* = 0.003, $p = .949$.

The participants improved from Block 1 (mean = .86, *SEM* = 0.01) to Block 2 (mean = .90, *SEM* = 0.01). Their accuracy then remained at a similar level in Block 3 (mean = .89, *SEM* = 0.01). The difference in accuracy was found to be statistically significant, $F_{(2, 86)} = 9.72$, *MSE* = 0.002, $p < .001$, $\eta_p^2 = .184$. Post hoc Sidak-corrected tests indicated that accuracy in Blocks 1 and 2 differed significantly ($p < .001$), as did Blocks 1 and 3 ($p = .006$). There was no difference in accuracy between Blocks 2 and 3 ($p = .855$).
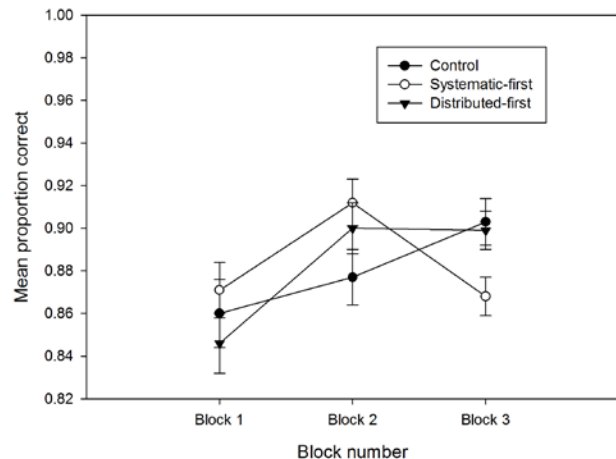
Figure 3: The interaction between condition and block.

There was a significant condition x block interaction, $F_{(4, 86)} = 3.59$, *MSE* = .002, $p = .009$, $\eta_p^2 = .143$. The interaction plot is displayed in Figure 3.

The interaction diagram suggested that participants in the Systematic-first condition may have been confused by having to switch from one systematic method to the other. It was thus considered more meaningful to focus on relative levels of improvement between Block 1 and Block 2. To this end, an improvement score was calculated by subtracting the Block 1 overall accuracy score from the Block 2 overall accuracy score. A one-way unrelated ANOVA was then conducted on this accuracy score. Whilst it did not reveal a statistically significant difference between the conditions, $F_{(2, 43)} = 1.36$, *MSE* = 0.004, $p = .268$, there was a greater improvement in the overall accuracy of label-checking for the Systematic (mean = 0.041, *SD* = 0.05) and the Distributed (mean = 0.054, *SD* = 0.07) conditions over the control condition (mean = 0.017, *SD* = 0.06).

In order to determine whether label errors were better detected after initial practice, performance in Block 2 was compared on trials containing a mismatch between the two sources of information. The participants in the Systematic condition (mean = .78, *SD* = 0.09) were the most accurate in detecting label errors, followed by the Control condition (mean = 0.71, *SD* = 0.12), and then the Distributed condition (mean = 0.64, *SD* = 0.20). A one-way unrelated ANOVA indicated that the difference in accuracy between the conditions was significant, $F_{(2, 43)} = 3.99$, $p = .026$, $\eta_p^2 = .157$). Post hoc Sidak-corrected tests indicated that the Systematic and Distributed conditions differed significantly ($p = .021$). There was no significant difference between the Systematic and Control conditions ($p = .496$) or the Control and Distributed conditions ($p = .500$).

DISCUSSION

The purpose of this study was to determine whether the quality control label-checking process could be improved by changing behaviour. Two computer-coupled methods were implemented, a systematic approach which visually guided participants and a distributed approach which "read out" the information to be checked, in addition to the visual guide. Overall, the study demonstrated improvements in checking labels over time, with performance in Block 2 being more accurate than in Block 1. Over-and-above that general improvement across conditions, Block 2 indicated a substantially larger improvement in the overall accuracy of responses under the conditions which imposed a systematic approach upon participants relative to the condition wherein participants continued to use their own self-chosen method of checking. However, in Block 3, whilst the overall accuracy of performance continued to rise for the control condition, it dropped substantially for the Systematic-first condition and plateaued for the Distributed-first condition.

When designing the experiment, a decision was taken to attempt to overcome potential individual differences in participants' error-checking abilities. Consequently, the participants in the two experimental groups performed both computer-coupled strategies in a counter-balanced order. Although the principles behind this decision were methodologically sound, a task-switching effect seems to have been introduced due to the change in task demands [13]. This task-switching effect can be seen in the reduction of overall accuracy for the distributed condition and a plateau for the systematic condition in Block 3. In order to interpret the data without the impact of these task-switching effects, the Discussion focuses hereafter on the results from Blocks 1 and 2 only.

In Block 1, all participants performed the label-checking task with no specific strategy requirements. For overall task accuracy (see Figure 3), this baseline level of performance was similar for the control participants compared with those participants that would go on to use a computer-led approach in Block 2. As expected, the performance of the control participants showed a slight improvement between Block 1 and Block 2. However, both computer-led strategies led to improved accuracy in Block 2 to a greater extent than the control participants (note the errors bars of the two strategy groups do not overlap with that of the control group). This is an encouraging result as it implies that the strategies implemented did adapt behaviour to design. This modest 5% improvement could relate to significant financial and environmental savings over time.

Further to the analyses performed on overall accuracy, accuracy in response to the mismatch trials only was also examined. It is important to note that, due to the nature of the task (with one in five trials being a mismatch trial; still significantly higher than would be experienced in the real-world setting), these analyses are low on statistical power - including variability in performances within-groups which may augment and/or mask any potential effects. Looking solely at accuracy for mismatch trials, in Block 2 those participants who used the systematic strategy had a higher degree of accuracy when detecting these mismatches than both the control participants and those participants who were instructed to use the distributed strategy. The unexpectedly lower accuracy for the distributed strategy is likely to be due to the low statistical power mentioned above (the variability in performance, measured in terms of SD, is twice that seen in those participants who used the systematic strategy, and 25% more than the controls).

From these data, it would appear that performance can be improved if the checking process were to be shared between a human operative and a computer system designed in such a way as to reduce attentional drift, the misdirection of perception, and cognitive load. Future research should seek to determine the extent to which specific components of cognition are prone to failure and are also, therefore, open to improvement.

CONCLUSION

Although checking tasks such as the one described in this paper are ostensibly very simple, they are likely to rely on multiple, interacting cognitive processes, any of which can be taxed beyond capacity and, as a result, cause accuracy to suffer. From the current findings, the adoption of a systematic approach to checking would appear to have reduced one or more factors which cognitive efficiency. At the very least, such approaches ensure that attention remains focused and that working memory functions optimally within its limits. Ensuring that both of these concerns are met is essential, especially in situations where sustained, repetitive checking of similar items is required [14]. Although the improvement in detecting errors reported in this paper may be modest in absolute terms (approximately 5%), every mistake detected in the context of the present investigation would amount to substantial financial and environmental savings by preventing large-scale food recalls and wastage.

The relevance of the current research goes beyond the fresh produce industry and indeed label-checking in general. There are many other situations in which detecting a mismatch between 'definitive' information and its re-presentation is critical. What is presented in one place or at one point in time needs to be transferred correctly to another place or point in time. If it is not, then this incorrect transfer needs to be detected with as high a degree of accuracy as possible.

There are numerous examples in the scientific literature of otherwise salient environmental features failing to be noticed as a result of fatigue, inattention, distraction, interference or excessive cognitive demand [15-19]. The consequences that may follow from similar errors in real world situations such as security checks, blood transfusions [19], pharmacological dispensing [20], air traffic control

[21] and driving [22] are only too obvious. Given the high costs involved, methods which minimize errors by ensuring the successful management of cognitive resources through task design and implementation become all the more important.

REFERENCES

1. Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 2 (1956), 81-97.

2. Baddeley, A. D., and Hitch, G. J. Working memory. In G. Bower (Ed.), *The psychology of learning and motivation* (pp. 47-89), Academic Press, New York, USA, 1974.

3. Baddeley, A. D. *Working memory.* Clarendon Press, Oxford, UK, 1986.

4. Zhang, J. The nature of external representations in problem solving. *Cognitive Science, 21,* 2 (1997), 179-217.

5. Zhang, J., and Norman, D. A. Representations in distributed cognitive tasks. *Cognitive Science, 18,* 1 (1994), 87-122.

6. Scaife, M., and Rogers, Y. External cognition: How do graphical representations work? *International Journal of Human-Computer Studies 45,* 2 (1995), 185-213.

7. Glasspool, D. W., Oettinger, A., Smith-Spark, J. H., Fox, J., Castillo, F. C., and Monaghan, V. E. L. Supporting medical planning by mitigating cognitive load. *Methods of Information in Medicine, 46,* 6 (2007), 636-640.

8. Hutchins, E. *Cognition in the wild*, Cambridge, MA, USA: MIT Press, 1995.

9. Clark, A., and Chalmers, D. J. The extended mind. *Analysis, 58,* 1 (1998), 10-23.

10. Kirsh, D., and Maglio, P. On distinguishing epistemic from pragmatic action. *Cognitive Science, 18,* 4 (1994), 513-549.

11. Sweller, J. Cognitive load during problem-solving: Effects on learning. *Cognitive Science, 12,* 2 (1988), 257-285.

12. Boksem, M. A. S., Meijman, T. F., and Lorist, M. M. Effects of mental fatigue on attention: an ERP study. *Cognitive Brain Research, 25, 1* (2005), 107-116.

13. Monsell, S. Task switching. *Trends in Cognitive Sciences, 7,* 3 (2003), 134-140.

14. Paus, T., Zatorre, R. J., Hofle, N., Caramanos, Z., Gotman, J., Petrides, M., and Evans, A. C. Time-related changes in neural systems underlying attention and arousal during the performance of an auditory vigilance task. *Journal of Cognitive Neuroscience, 9,* 3 (1997), 392-408.

15. Robinson, A., Manzi, A., and Triesch, J. Object perception is selectively slowed by a visually similar working memory load. *Journal of Vision, 8,* 16 (2008), 1-13.

16. Mack, A. Inattentional blindness: Looking without seeing. *Current Directions in Psychological Science, 12,* 5 (2003), 180-184.

17. Lustig, C., May, C. P., and Hasher, L. Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General, 130,* 2 (2001), 199-207.

18. Hsu, K. E., Man, F-Y., Gizicki, R. A., Feldman, L. S., and Fried, G. M. Experienced surgeons can do more than one thing at a time: Effect of distraction on performance of a simple laparoscopic and cognitive task by experienced and novice surgeons. *Surgical Endoscopy, 22,* 1 (2008), 196-201.

19. Sharma, R. R., Kumar, S., and Agnihotri, S. K. Sources of preventable errors related to transfusion. *Vox Sanguinis, 81,* 1 (2001), 37-41.

20. Ross, L. M., Wallace, J., and Paton, J. Y. Medication errors in a paediatric teaching hospital in the UK: Five years operational experience. *Archives of Disease in Childhood, 83,* 6 (2000), 492-497.

21. Moon, W.-C., Yoo, K.-E., and Choi, Y.-C. Air traffic volume and air traffic control human errors. *Journal of Transportation Technologies, 1,* 3 (2011), 47-53.

22. Stanton, N. A., and Salmon P.M. Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems. *Safety Science, 47,* 2 (2009), 227-237.