The ability to interpret mathematical and statistical information at a level that is functional in today's information-laden society is, for many individuals, an essential skill. Figures and statistics are routinely presented via newspapers, television, and the internet, and often in a way intended to influence how the information is received (Tufte, 1983). Nevertheless, the accurate interpretation of this information is vital in making informed choices and decisions, especially within the workplace (Huff, 1993). Despite the importance of such skills, studies have shown that the general level of numeracy in the UK is poor. For instance, the results from a survey carried out in England found that approximately 15 million adults had only entry level numeracy skills (Grinyer, 2005), and results from a survey conducted for the Welsh Assembly Government (2005) illustrated that 53% of adults did not have level one numeracy skills. A concerning fact when you consider that in the UK students take General Certificates of Secondary Education (GCSE) examinations when they are 16 years old and level one numeracy equates to the lowest achievable grade mark without failing (i.e., approximately below the GCSE's G grade standard).

There is some evidence that the level of numeracy skills in new psychology undergraduate students is deteriorating. A recent study...
comparing cohorts who entered psychology departments a decade apart highlighted declining standards in mathematical abilities (Mulhern & Wylie, 2004), and striking deficiencies in mathematical reasoning have also been identified when comparing across a sample of UK educational institutions (Gnaldi, 2006; Mulhern & Wylie, 2005). Undergraduate students taking science degrees face an even greater requirement to become familiar with data interpretation and to become competent with statistics, so that they are able to use systematic, objective methods for summarising and interpreting quantitative information (Everson, Zieffler, & Garfield, 2008). Students must be able to understand statistics so that they can conduct their own research effectively and interpret and evaluate the research of others. Therefore, students need to become fluent in the vocabulary of basic statistical terms and concepts (Lalonde & Gardner, 1993); something that many students find difficult. The complexities involved in learning about statistics have been likened to those associated with learning a second language (Lalonde & Gardner, 1993; Lazar, 1990).

When learning to play a musical instrument, when training to become a professional athlete, or when learning a second language, repeated practice is the accepted route to an accomplished performance (Binder, 1996; Kubina & Morrison, 2000). Furthermore, a learner may reach a level of mastery that distinguishes their performance from that of an average learner: a fluent performance can be characterised as effortless, almost automatic, accurate responding (Binder, 1996). This type of performance has also been referred to as automaticity (Bloom, 1986).

Fluency in basic skills, or elements, may be the basis of what we refer to as ‘mastery’ levels of performance and may be the prerequisite for learning higher-order more complex skills, or compounds (Kubina & Morrison, 2000). There are recognised sequences in which to learn each skill element. For example, each step of a sequence in conducting mathematical operations should be mastered before proceeding onto the next (Stein, Silbert, & Carnine, 1997). Problems may occur when a student lacks such mastery in the necessary skill elements that are required for fluent performance on the next level of that skill (Binder, 1996; Kubina & Morrison, 2000). In these circumstances, a learner may be thought not to ‘understand’ the concept being presented. However, this can better be conceptualised as the learner being dysfluent at skill elements of the compound task (Binder, 1993, 2003; Binder, Haughton, & Bateman, 2002; Hughes, Beverley, & Whitehead, 2007; McDowell & Keenan, 2001).

A deficit in skill elements prevents learners from further progression through a curriculum, a term that Binder (1996) referred to as cumulative dysfluency. To ensure that a skill is mastered before moving on, specific performance targets or fluency aims must be set that allow the learner to demonstrate mastery on the existing skill element prior to moving on to the next. Detailed fluency aim ranges of appropriate rates of accurate responding have been researched and documented for many academic and motor skills (Kubina, 2002). For example, to read effectively a reader should be able to read aloud between 200-250 words of prose per minute and failure to be able to read at this rate may mean that the very process of reading is difficult and aversive to a learner (Hughes, et al., 2007). When the specific aims of fluent performance are achieved, there are benefits that result from this level of performance. The outcomes of fluent performance have been captured in the acronym RESA: Retention, Endurance, Stability, and Application (Fabrizio, 2004; Fabrizio & Moors, 2003; Johnson & Layng, 1992). Retention is the ability to perform the skill fluently after long periods without practice, endurance is the ability to perform the skill for longer periods of time, stability is the ability to perform the skill in the presence of other distractions, and application is the ability to combine previously learnt element sub-skills to perform novel, more complex, compound skills (Kubina & Morrison, 2000).

Statistics teachers face many problems in teaching such a difficult subject area. According to Conners, McCown, and Roskos-Ewoldsen (1998) there are four main problems: variability in the performance of individuals, student
motivation, anxiety about statistics, and making learning last. Applied behaviour analytic approaches to teaching have a long history of success in overcoming these and other obstacles to learning (Gardner, et al., 1994; Heward, 2005; Heward, et al., 2005; Johnson & Layng, 1994; Layng, Twyman, & Stikeleather, 2003; Moran & Malott, 2004). As Frederick and Hummel (2004) comment, effective educational practices—those that produce good educational outcomes—adhere to certain principles, all of which are relevant to these four challenges:

Effective instruction begins with clearly stated behavioral objectives; provides accurate, competent models; provides many opportunities for active responding; delivers immediate feedback about the accuracy of responses; allows self pacing; teaches to mastery; reinforces accurate responding; and frequently and directly measures responding that is explicitly tied to behavioral objectives, using the outcomes of those measurements to make instructional decisions. (p. 11)

Precision Teaching (PT) has proved to be an effective intervention that can monitor the effectiveness of any teaching method and also evaluate the outcomes of learning—RESA. PT has over four decades of evidence supporting its use (Beck & Clement, 1991); and in terms of meeting the eight principles of effective instruction (Frederick & Hummel, 2004), PT and frequency-building procedures do extremely well (Beck & Clement, 1991; Binder, 1988; Binder, et al., 2002; Chiesa & Robertson, 2000; Hughes, et al., 2007; Kubina & Morrison, 2000; McDade, 2005). PT procedures ensure that all skills have clearly stated behavioural objectives, provide many opportunities for responding, often give immediate feedback about the accuracy of each response, and allow learners to work at their own pace until they meet predefined levels of performance—fluency aims. Learners who follow PT and fluency-based procedures often use short, timed practice sprints to collect data on their performance. These timed practice sessions are usually carried out daily. They record the number of correct and incorrect responses per minute (Learning Opportunities; LO) and then plot these data on a Standard Celeration Chart (SCC). The SCC and its use have been described in detail in a number of previous publications (e.g., Calkin, 2005; Kubina & Yurich, in press; West, Young, & Spooner, 1990; White, 1986; White & Neely, 2004) but essentially its most obvious use allows both the learner and teacher to easily observe whether effective learning is taking place. After only a few days of plotting data, a SCC produces a learning picture that enables data-driven instructional decisions to be made in order to improve an individual’s learning (Claypool-Frey, 2009a; Lindsley, 1995; McGreevy, 1983).

SAFMEDS (Say All Fast Minute Every Day Shuffled) are a practice and assessment procedure developed to help students learn and build fluency on key facts (Graf & Lindsley, 2002). SAFMEDS are typically used to help students become fluent in definitions and basic concepts, thus making complex learning and later synthesis of concepts more likely. Further details of SAFMEDS usage have been described in earlier publications (Claypool-Frey, 2009b; Graf & Lindsley, 2002; Vieitez, 2003).

In the current study we explored the effects of a brief intervention using SAFMEDS cards within a PT and frequency-building framework designed to build high frequency performance in key statistical definitions and concepts in a group of first year psychology undergraduates. We examined whether student performance on basic statistics facts (elements) would influence their performance on the post-test (compound).

Methods

Participants

Initially we sought informed consent from an entire first year statistics class containing approximately 340 students. From those who gave consent, we identified those who scored at the 50th percentile and below as determined by their first weekly class test scores, and invited them to sit a pre-test that consisted of questions selected to cover all the statistical content for the semester. We randomly allocated students to either the Precision Teaching (PT) or Treatment as Usual (TAU) group using a stratified random sampling procedure (Gravetter & Forzano, 2009): to ensure that the groups were balanced
we allocated students to blocks of equally performing students and randomly assigned from each of those blocks to the two groups (each block contained an even number from which to randomly assign). Participants’ ages ranged from 18-30. Initially the participants \((n = 66)\) were split into two equal size groups, however by the end of the study attrition rates had reduced the number of participants \((n = 55)\), to 24 in the PT group and 31 in the TAU group. Participants earned course credits for taking part in the study.

**Stimuli & Materials**

We constructed a pre-test that comprised 50 randomly selected questions from the core text’s bank of test questions (Gravetter & Wallnau, 2007). The tests consisted of multiple-choice questions (MCQ) from all the content text-book chapters to be covered that semester. We prepared two packs of SAFMEDS cards (each containing 80 cards) for each participant in the PT group. The packs covered course content for the entire semester (Pack 1: weeks 3-5 of the course & Pack 2: weeks 6-11) and were based on materials taken from the core text’s study guide (Gravetter & Wallnau, 2004). We also gave participants data collection sheets with a 10-week 3-cycle Standard Celeration Chart (SCC) for charting their learning, and digital timers to use during practice sprints and timings.

**Design**

For the purpose of group data analysis, we used a mixed design with one between-group factor (Group: PT & TAU) and one repeated measures factor (Test: pre- & post-test) to allow the comparison of group performance between the pre- and post-tests.

As is typical when using behavioural research methodologies, we examined each participant’s ongoing graphed data to make intervention decisions on an individual basis.

**Procedure**

All students who had been invited to participate in the study took the initial 50-item pre-test and were given 45 minutes to complete. Following the initial pre-test, students who were allocated to the TAU group continued with their standard lectures and their normal preparation for their weekly open-book test. Students who were allocated to the PT group attended lectures as normal but were also instructed in the use of SAFMEDS and how to chart learning performance. Students from both the PT and the TAU group had the opportunity to attend a two-hour lecture and a one-hour revision session each week. However, approximately 30 minutes of each 2-hour lecture were used to administer the weekly class tests from week 3 onwards. Students in both groups received the same tests under the same conditions within the class setting. The same lecturer presented all lectures and revision sessions on the course.

**Precision teaching procedure.** We held an initial one-hour session for the entire PT group during which we tutored them in PT methods, the rationale for our approach, and provided them with examples of the research evidence that supports the use of PT methods. During the session we emphasised the importance of data collection, using the SCC to monitor learning progress, and how to effectively use SAFMEDS cards (see below). The PT group was split into subgroups each with their own proctor—an undergraduate student helping with the research. Participants were instructed to conduct 1-minute timings three times per day, collect data for each timing, and plot their best score for each day onto their SCC. This would take them approximately six minutes per day. We also asked them to meet twice weekly in their subgroup, so that their proctor could verify their SAFMEDS and charting procedure and help them decide on strategies that may aid their learning. Whilst they were involved in the PT intervention they continued to take their open-book weekly tests.

At the end of the semester, participants from both groups completed a post-test; this post-test was the same test they had previously taken as a pre-test.

**SAFMEDS procedures.** We constructed our packs so that there was a statistical definition on the front of each card with a key term of that definition missing. On the reverse of the card, the missing term was printed (see Figure 1 for
example of card content). This then required the learner to see the definition on the front of the card, and then to say aloud the missing term that was printed on the reverse of the card. They then would turn the card over for immediate feedback and place the cards onto the table in one of two piles—on the right pile if the answer was correct and on the left if it was not. Following each 1-minute timing, participants would count up the number of correct and incorrect responses (learning opportunities) and record these scores on their data sheet. Their best scores for each day would then be plotted on the SCC. Although most of the SAFMEDS had terms and definitions, some had a statistical symbol on the front with the name or the symbol’s interpretation on the reverse. None of the SAFMEDS cards was based on questions from the pre- and post-test, only key terms and definitions.

Initially we issued each participant with the first pack of SAFMEDS. This pack covered terms and definitions from the chapters that they had covered in weeks 3 and 4, and that they would be tested on in the following week’s test (week 5). After the test in week 5, we gave them the second pack designed to cover content from chapters for the rest of the course. From this point they would conduct timings for both of the packs working towards the pre-determined fluency aim range, and continue either until the end of the study or when they reached aim with either or both packs.

If participants reached an aim that was considered fluent with a pack, they continued working with the other pack whilst reviewing their mastered pack at weekly intervals to ensure retention of content. The fluency aim for the SAFMEDS cards was 60–40 correct with no more than 2 errors per minute (Kubina, 2002).

Interventions. Examples of possible interventions were: changes in timing period or number of timings conducted, slice-backs in number of cards worked with, and introduction of error correction procedures. For the error cor-

<table>
<thead>
<tr>
<th>Front of Card</th>
<th>Back of Card</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard deviation is the _____</td>
<td>square root</td>
</tr>
<tr>
<td>_____ of the variance</td>
<td></td>
</tr>
<tr>
<td>bimodal</td>
<td>____ distribution</td>
</tr>
</tbody>
</table>

Figure 1. Example of the front and back of two of the SAFMEDS cards from Pack 1.
rection procedure, for example, we instructed the students that before conducting their daily timings they should first conduct an untimed pack review. During this process they would review each card in the pack, as they would do with a normal timing, placing corrects in one pile and learning opportunities in another. Then they would take the pile of learning opportunities and review these again following the same procedure. They would continue to repeat this error correction procedure until they had said out loud one correct response for every card in their pack. They would then conduct their daily timings as usual and plot their best score of the day.

Results

An ANCOVA (controlling for any differences between the groups at pre-test) was conducted on the data to compare the mean percentage correct scores at post-test. This showed a significant difference in favour of the PT group with a medium effect size, $F(1, 53) = 5.23, p = .026, d = 0.62$ (see Figure 2).

Figure 3 shows the performance of the PT and TAU groups on each of the weekly open-book tests taken in weeks 5, 6, 7, 9, 10, and 11. The PT group consistently had higher mean percentage scores for all of the weekly tests. Figure 3 clearly shows that the PT group’s mean score was over 10% higher than that of the TAU group for the first test after beginning the SAFM Edwards (week 5). This was after only one week of working with the first pack of cards.

Figure 4 shows the percentage improvement from pre- to post-test scores for the PT and the TAU groups at a more individual level. It can be seen that almost all participants in the PT group improved on their score from pre- to post-test with no students showing a deterioration in performance (improvement range = 0 to 42); whilst in the TAU group six participants made either no gain or showed a deterioration in performance between pre- and post-test (improvement range = -8 to 28). Eight participants (33.33%) from the PT group increased their score by more than 2 standard deviations, whereas only 5 (16%) of the TAU group did so. Furthermore, from these 8 PT participants 4 increased their scores at or above 3 standard deviations.

Figure 2. Pre- and post-test mean percentage scores for the PT and TAU groups. Error bars represent ± 1 standard error of the mean. Covariates appearing in the model (pre-test) were evaluated as 41.20.
Figure 5 illustrates the use and interpretation of a SCC at the individual level for one participant from the PT group. This learner began conducting six timings a day working with Pack 1, and recording their best daily score, after they had already taken the week four statistics test.

As can be seen from Figure 5 there are no gaps in the data with the participant even conducting timing sprints over the weekend.

During the period of working solely with Pack 1 the charted data displays a classic cross-over jaws learning picture, at first errors (LO)
are high and corrects are low, however after four days of practice timings, the errors begin to decrease as the corrects increase.

When the learner begins to work with both packs there is an apparent increase in LO for Pack 1 as they jump up above corrects again. However the LO soon begin to decrease again over each subsequent week. The first timing for Pack 1 was 3 correct and 10 LO (13 cards attempted in one minute) and for the first 2 weeks the learner averaged approximately 15 cards worked through in a 1-minute timing.

For the following two weeks the learner was working with both packs (conducting 3 timings with each pack). The first timing with Pack 2 was 4 correct and 19 LO (23 cards attempted in one minute), and the average cards worked through in a 1-minute timing were approximately 31 and 28 for packs one and two respectively.

Discussion

The aim of this study was to evaluate the effectiveness of a brief intervention using SAFMEDS cards within a framework of PT,
and frequency-building procedures designed to build mastery performance in key statistical definitions and concepts in a group of first-year psychology undergraduates. The students selected had scored at the 50th percentile and below at their initial weekly statistics test, so were those students who were performing below the average.

There was no statistically significant difference when comparing mean test scores between the two groups at pre-test. However, at post-test (after controlling for pre-test score differences) the mean score was significantly higher for the PT group in comparison to the TAU. As expected both groups improved during the semester, however, the PT group improved by almost one entire grade level above the performance of the TAU group (i.e., from a C- to a B-); the categorical marking system is a letter grade system ranging from F (Fail) to A (Distinction), with each of these having three sub-categories (e.g., B-, B, B+).

When comparing group performance on the weekly tests (weeks 5, 6, 7, 9, 10, & 11) the trend was consistently higher for the PT group (see Figure 3). The difference between the two groups was more noticeable for the first test that participants took after the initial week of practicing with the first pack of SAFMEDS (week 5). The PT group outperformed the TAU group by an average of over 10% for that week, which provides additional support for the PT group having become more versed in the basic definitions and terminologies. Recall that the test in week five was still concerned with the early chapters of the core text and success in this test relied heavily on students having mastered the earlier taught content.

Although we did not collect data to validate the class tests that were used as pre- and post-test measures, they are clearly statistic questions written by respected authors whose book is now in the seventh edition and used as a required text in many psychology departments. Furthermore the instructor test bank items were based on the curriculum, all students initially performed poorly at baseline, and the same test was used for both pre- and post-test measures.

When comparing pre- to post-test improvement at an individual level it was found that twice as many from the PT group increased their score by over two standard deviations in comparison to the TAU group, and half of these achieved or bettered an increase of three standard deviations (see figure 4). We would regard this type of improvement as statistically and educationally significant.

Considering the example SCC for one individual presented in the results, it can be seen that throughout this brief intervention the participant was consistent in conducting regular timings and recording their best daily score for each pack. Their corrects continued to increase and their LO decreased, although they did not quite reach the set fluency aims. At the point that the second pack was introduced the LO seem to take a jump up (rising above the number of corrects) for Pack 1; although at this same point in time for Pack 2 there is a classic picture of beginning learning in that LO are high and corrects are low. This then progresses into a crossover jaws learning picture as the other pack did.

The explanation for the jump up for Pack 1 would be that although the LO did increase temporarily, this was essentially because the learner had become more practiced and comfortable with the procedure and increased the number of cards they attempted during sessions. This is something that would not show up so clearly if using the traditional percent correct measure of academic success. The learning picture for Pack 1 for the first two weeks is a typical picture of a learner who is concentrating on achieving correct answers, irrespective of how long it takes them to think of the response. During this period the learner’s average response rate is less than 15 responses per minute (1 every four seconds). Following this initial two-week period the average response rate was approximately 30 cards per minute (1 response every 2 seconds) for both packs.

Some issues of the research methodologies merit discussion. In the present design, the daily use of SAFMEDS could not be monitored with complete accuracy; we had to rely on self-report of the PT participants at their bi-weekly meetings; although we did have participants conduct
timings at each meeting (in the presence of their student proctor) to ensure their recorded data matched their performance data. Neither were data collected to monitor attendance at weekly lectures or the time spent by each participant in study and preparation for the weekly tests.

In addition, because the first test in week 3 was used to identify participants for possible inclusion in the study, students did not receive their first pack of cards until after the test in week 4; therefore they only had one week in which to work with the first pack until they sat the next test in week 5, following which they received their second pack of SAFMEDS cards. It would have been preferred to pre-test the entire year and then invite participants to take part in the study, so that they could aim for a fluent performance in the first pack before sitting their first test in week 3. However, this would demand considerably more resources than we had at our disposal.

Whilst all possible control checks were followed to ensure consistency in the definitions and concepts in the SAFMEDS packs, there were some perhaps unavoidable differences between some of the cards. As we included statistical symbols for some cards and more complex definitions and concepts for others it is likely that different cards could be responded to at different latencies; this could have had an adverse effect on an individual's ability to perform equally well on all cards in a pack. Nevertheless in considering the SCC for this individual there is very little variation in their performance from day to day.

Further research would benefit from some changes to the methodology to overcome some of the problems we encountered during this study. The students in the PT group had incentives to study that were not available to the TAU group. Future research should ensure that a comparison group also receive instruction on using key-term flashcards but no instruction on how to time or chart their performance. Additionally, the results found could be more robustly attributed to the effects of the SAFMEDS intervention if this comparison group were to not only revise using standard methods, but also meet up with a student proctor bi-weekly (as the PT group did in this intervention).

The advantages in conducting frequency-based interventions via the internet are many. We could have SAFMEDS presented via an internet-based system, that would allow easy tracking of effective engagement of the participants, tracking both the number of timings conducted and also each of the timings scores, as well as automatically charting and displaying their progress. The programme could be designed to be even more adaptive to each individual learner, dropping cards temporarily from their pack when the participant had mastered them (i.e., consistently shows correct responding and low response latencies for that card) and readmitting them to the pack at set periods to ensure that retention is maintained. This would allow the more difficult cards for each learner to be presented more often. Finally, a web-based intervention would allow the automatic measurement of not just participant's rate of responding but also the measurement of response latency and the outcomes of fluency defined by RESA (i.e., Retention, Endurance, Stability, & Application) (Fabrizio, 2004; Fabrizio & Moors, 2003; Johnson & Layng, 1992). Of course by definition someone who is fluent would demonstrate low response latencies; it would be interesting to look at the relationship between response latencies and these outcomes of fluent responding (i.e., RESA) more systematically.

If we were to also add contingencies so that students would gain some credit towards the module assessments for participating, this could increase their motivation for active engagement with the procedure. Future research could also benefit from recording both groups attendance at the weekly statistic lectures, interviewing participants about their revision strategies, and measuring statistics anxiety to discover whether the PT and frequency-building procedures may reduce anxiety levels (Conners, et al., 1998).

This study set out to improve undergraduate statistic performance using a combination of PT and frequency-building procedures. These procedures included daily timed practice sprints using SAFMEDS cards, charting participant's best daily score on SCC, and using these plot-
ted data to make instructional decisions to help participants to progress towards their goal at an optimal rate. The study provides some additional support for the merits of such interventions to enhance learning that are effective, can be easily implemented, and are of relatively low cost. Such interventions have been successfully used previously (Beck & Clement, 1991; Chiesa & Robertson, 2000; Hughes, et al., 2007) and are suitable to be used with any age group and with almost any curriculum. Although in this study we targeted psychology undergraduates undertaking a statistics module, if more widely adopted, these techniques could help to diminish the skill deficits that a great number of both children and adults currently experience.

References


Gnaldi, M. (2006). The relationship between poor numerical abilities and subsequent difficulty in accumulating statistical knowl-


MO: Plain English Publications (University of Missouri).