Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

To request permissions please use the Feedback form on our webpage.
http://researchspace.auckland.ac.nz/feedback

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form and Deposit Licence.
Representativeness of Direct Observations of Behavioural Duration:

*Data-based procedures for selecting a sampling method*

REBECCA ANNE SHARP

A thesis submitted in fulfilment of the requirements of the degree of Doctor of Philosophy in Psychology, School of Psychology, The University of Auckland, 2014.
<table>
<thead>
<tr>
<th>CHAPTER 1: REPRESENTATIVENESS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINITION OF TERMS</td>
<td>3</td>
</tr>
<tr>
<td>MEASUREMENT OF REPRESENTATIVENESS</td>
<td>8</td>
</tr>
<tr>
<td>FACTORS AFFECTING REPRESENTATIVENESS</td>
<td>10</td>
</tr>
<tr>
<td>SAMPLING METHOD</td>
<td>11</td>
</tr>
<tr>
<td>DIMENSIONS OF BEHAVIOUR</td>
<td>26</td>
</tr>
<tr>
<td>SELECTING SAMPLING METHODS</td>
<td>31</td>
</tr>
<tr>
<td>SUMMARY</td>
<td>33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 2: WORK SAMPLING</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CASE EXAMPLE</td>
<td>35</td>
</tr>
<tr>
<td>BACKGROUND</td>
<td>37</td>
</tr>
<tr>
<td>SELECTION OF BEHAVIOURAL CATEGORIES</td>
<td>41</td>
</tr>
<tr>
<td>DEFINITION OF BEHAVIOURAL CATEGORIES</td>
<td>42</td>
</tr>
<tr>
<td>DATA COLLECTION</td>
<td>43</td>
</tr>
<tr>
<td>WORK SAMPLING EQUATION</td>
<td>49</td>
</tr>
<tr>
<td>NOMOGRAPHS</td>
<td>58</td>
</tr>
<tr>
<td>CONTROL CHARTS</td>
<td>62</td>
</tr>
<tr>
<td>INITIAL ESTIMATES OF $p$</td>
<td>67</td>
</tr>
<tr>
<td>SELECTING OBSERVATION TIMES</td>
<td>68</td>
</tr>
<tr>
<td>DATA ANALYSIS</td>
<td>77</td>
</tr>
<tr>
<td>ADVANTAGES AND LIMITATIONS</td>
<td>79</td>
</tr>
<tr>
<td>EXAMPLES OF WORK SAMPLING STUDIES</td>
<td>84</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 3: EVALUATION OF WORK SAMPLING METHODS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>STUDY 1 RATIONALE</td>
<td>95</td>
</tr>
<tr>
<td>METHOD</td>
<td>103</td>
</tr>
</tbody>
</table>
OVERALL DURATION ................................................................. 194
BOUT DURATION ................................................................. 196
LOW-DURATION BEHAVIOURS ..................................................... 198
MOMENTARY TIME SAMPLE INTERVAL DURATION ......................... 202
LONGER BOUT DURATIONS ............................................................ 202
3-D MESH PLOTS .................................................................. 205
3-D STEM PLOTS ................................................................ 212
ANALYSIS OF DECISION RULES .................................................. 215
SELECTION OF STUDY 1 BEHAVIOURS ........................................ 215
SAMPLING ........................................................................ 217
DISCUSSION ........................................................................... 221
DIMENSIONS OF BEHAVIOUR ..................................................... 222
SAMPLING METHOD PARAMETERS ................................................ 231
LIMITATIONS ........................................................................ 234
STRENGTHS .......................................................................... 238
DATA-BASED DECISION RULES .................................................... 239
CONCLUSION .......................................................................... 240

CHAPTER 6: GENERAL DISCUSSION

DETERMINING REPRESENTATIVENESS ............................................ 243
UTILITY OF DATA-BASED METHODS ............................................. 243
LIMITATIONS OF ASSUMPTIONS .................................................... 244
PRACTICALITY OF DATA-BASED METHODS ..................................... 245
SUGGESTIONS FOR FUTURE RESEARCH ........................................ 246
CONCLUSION .......................................................................... 248

GLOSSARY ......................................................................... 250
REFERENCES ...................................................................... 253
APPENDICES

1: TEACHER ESTIMATES ............................................................. 281
2: DURATION ANALYSIS ............................................................ 285
3: MATLAB® CODE ................................................................. 294
4: ROTATED 3-D MESH PLOTS & 3-D STEM PLOTS ......................... 299
Behaviour analysts have sought to develop effective methods for selecting observation methods that reflect overall dimensions of behaviour across the whole time-of-interest. The existing literature on representativeness provides general rules for selecting representative measurement systems, however applied behaviour analysis would benefit from a data-based method. My research evaluated two data-based methods. In Study 1, the utility of work sampling, a method from time-and-motion study used to determine efficiently how people spend their time in work settings, was tested. Full week observations of behavioural categories in children with developmental disabilities were conducted in a special school. Equations used in work sampling were used to select the number of observations to be extracted from criterion records to obtain representative samples. The data showed that the number of samples required was impractically high when relative accuracy was the independent variable, and that representative samples were not obtained when absolute accuracy was computed. In Study 2, computer simulations evaluated the effect of varying overall duration and bout duration on the representativeness of samples extracted using simulated momentary time sampling. Decision rules were developed in the form of 3-D graphs, from which practitioners are able to select a measurement system based on estimates of the dimensions of the behaviour (bout duration and overall duration) and acceptable error. The results showed that decision rules could only be developed for a limited range of overall durations of behaviour. When applied to whole-week datasets obtained in Study 1, the decision rules did not produce representative samples when interbout intervals were variable. I conclude that, as the parameters of all dimensions of behaviour need to be known in order to predict representative sampling, a search for algorithms to predict representative sampling based on the parameters of some dimensions is a difficult task.
ACKNOWLEDGEMENTS

First, I would like to acknowledge Dr. Oliver Mudford for his guidance and teaching me the wide range of academic skills I now possess and hope continue to develop. Never was a request for help unanswered, from big discussions of difficult concepts to small 'stupidity checks' of my conference presentations. It is truly appreciated, thank you.

Second, I would like to acknowledge Dr. Douglas Elliffe, under whose guidance I have been for a number of years, and whose encouragement of my pursuit of achievement has never faltered. The number of skills I have learned from you is immeasurable. Thank you.

I am appreciative of my colleagues, Dr. Angela Arnold-Saritepe and Katrina Phillips, who always reinforced my mands for information, help, and attention. Thank you for your advice, and for all the occasions in which you engaged in listener behaviours for me.

I am appreciative of the assistance of numerous administrative and IT staff, who helped me with countless tasks including finding phone numbers, fixing printers, and loading programs onto computers. I also thank the special school in which I collected the data for my first study. The support of the communities in which behaviour analysts work and conduct research is vital to success and appreciated.

Finally, thank you to my family. My grumpiness never punished your attempts to ask me how my thesis was going, and you have been supportive over many years of academic study. Behind my PhD thesis are countless words of encouragement, home-cooked meals, hugs, and unconditional support from my family. Thank you.
Co-Authorship Form

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. Please include one copy of this form for each co-authored work. Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Approximately 5,400 words and two figures from Chapter 3. Submitted as:


Nature of contribution by PhD candidate: Student developed the study, obtained UAHPEC approval, recruited participants, collected the data, analysed the results and wrote the paper with comments from supervisors. The comments from supervisors weren’t significantly greater than would be expected on a chapter of a PhD thesis not directly intended for publication.

Extent of contribution by PhD candidate (%): 90%

CO-AUTHORS

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. Oliver Mudford</td>
<td>Advised on study through regular PhD supervision sessions. Read and provided feedback</td>
</tr>
<tr>
<td>Dr. Douglas Elliffe</td>
<td>Provided feedback on revised draft of manuscript before submission for publication.</td>
</tr>
</tbody>
</table>

Certification by Co-Authors

The undersigned hereby certify that:

☐ the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and

☐ in cases where the PhD candidate was the lead author of the work that the candidate wrote the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. Oliver Mudford</td>
<td></td>
<td>29/09/2014</td>
</tr>
<tr>
<td>Dr. Douglas Elliffe</td>
<td></td>
<td>30/09/2014</td>
</tr>
</tbody>
</table>

Last updated: 25 March 2013
This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. Please include one copy of this form for each co-authored work. Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

**In thesis as Chapter 5. Submitted as:**

**Nature of contribution by PhD candidate**
Ms Sharp conducted the study entirely. Also wrote 1st draft independently.

**Extent of contribution by PhD candidate (%)**
95%

## CO-AUTHORS

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. Oliver Mudford</td>
<td>Provided feedback on two drafts before the manuscript was submitted.</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Certification by Co-Authors

The undersigned hereby certify that:
- the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
- in cases where the PhD candidate was the lead author of the work that the candidate wrote the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. Oliver Mudford</td>
<td>[Signature]</td>
<td>30/09/2014</td>
</tr>
</tbody>
</table>
Co-Authorship Form

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. Please include one copy of this form for each co-authored work. Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Approximately 5,900 words and five figures from Chapter 5. Submitted as:


Nature of contribution by PhD candidate: designed and conducted study. Analysed and developed data-presentation. Wrote submitted ms.

Extent of contribution by PhD candidate (%): 90%

CO-AUTHORS

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. Oliver Mudford</td>
<td>Discussion about design, conduct, and data-presentation so I knew what she was doing. Read and commented on 1 draft of manuscript</td>
</tr>
<tr>
<td>Dr. Douglas Elliffe</td>
<td>Occasional discussion about PhD progress in general - not specific to manuscript</td>
</tr>
</tbody>
</table>

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- in cases where the PhD candidate was the lead author of the work that the candidate wrote the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. Oliver Mudford</td>
<td></td>
<td>19/12/2014</td>
</tr>
<tr>
<td>Dr. Douglas Elliffe</td>
<td></td>
<td>19/12/2014</td>
</tr>
</tbody>
</table>

Last updated: 25 March 2013
LIST OF FIGURES

FIGURE 1. Factors affecting the representativeness of samples of behaviour………… 6

FIGURE 2. Illustration of factors affecting representativeness…………………………. 7

FIGURE 3. The number of observations required by Equation 3 for 1% to 5% relative accuracy ($S$), as a function of the estimate of the proportion of time spent in the activity of interest ($p$)…………………………………………………….. 52

FIGURE 4. The number of observations required by Equation 4 for 1% to 5% absolute accuracy ($S$), as a function of the estimate of the proportion of time spent in the activity of interest ($p$)…………………………………………………….. 53

FIGURE 5. Nomograph showing the number of required observations calculated (Equation 4) from a $p$ value of 0.8 or 0.2, and required accuracy of 4% for a confidence level of 95%………………………………………………………… 58

FIGURE 6. An example control chart with constant control limits (three-sigma) constructed from theoretical data……………………………………………………. 64

FIGURE 7. An example control chart with changing control limits (three-sigma) constructed from theoretical data……………………………………………………. 65

FIGURE 8. The scheduling of observations under three momentary time sampling methods……………………………………………………………………. 71

FIGURE 9. Flowchart showing each of the analyses conducted………………………… 110

FIGURE 10. Representativeness of extracted momentary time samples (fixed, random, and discontinuous) shown as the obtained percentage of time spent in each behaviour ($P_0$) plotted against the overall daily percentage of time spent in that behaviour ($P_D$)…………………………………………………. 118

FIGURE 11. Representativeness of extracted continuous samples shown as the obtained percentage of time spent in each behaviour ($P_0$) plotted against the overall daily percentage of time spent in that behaviour ($P_D$)……………………………. 120
FIGURE 12. Effort (number of observations in each extracted observation session) plotted against the difference between the obtained percentage of time spent in a behaviour ($P_O$) and the overall daily percentage of time spent in that behaviour ($P_D$) for momentary time samples…………………………….. 123

FIGURE 13. Effort (number of observations in each extracted observation session) plotted against the difference between the obtained percentage of time spent in a behaviour ($P_O$) and the overall daily percentage of time spent in that behaviour ($P_D$) for continuous samples…………………………………….. 125

FIGURE 14. Effort (number of observations in each extracted observation session) plotted against the difference between the obtained percentage of time spent in a behaviour ($P_O$) and the overall daily percentage of time spent in that behaviour ($P_D$) for the three schedules of momentary time samples and continuous samples…………………………………………………. 127

FIGURE 15. Representativeness of extracted continuous samples commencing at differently scheduled times shown as the obtained percentage of time spent in each behaviour ($P_O$) plotted against the overall daily percentage of time spent in that behaviour ($P_D$)……………………………………………………….. 128

FIGURE 16. Effort (number of observations in each extracted observation session) plotted against the difference between the obtained percentage of time spent in a behaviour ($P_O$) and the overall daily percentage of time spent in that behaviour ($P_D$) for continuous samples…………………………………………….. 129

FIGURE 17. The difference between the weekly percentage of time spent in a behaviour ($P_W$) and the estimated percentage of time spent in that behaviour ($P_E$) as determined by teacher estimations or preliminary observations………………. 130

FIGURE 18. Representativeness of continuous samples extracted at times nominated by the teachers shown as the obtained percentage of time spent in each behaviour ($P_O$) plotted against the overall daily percentage of time spent in that behaviour ($P_D$)…………………………………………………………………. 131

FIGURE 19. Occurrence graph of on-task behaviour for Participant 1…………………………….. 132

FIGURE 20. Occurrence graph of on-task behaviour for Participant 2…………………………….. 132
FIGURE 21. Occurrence graph of running behaviour for Participant 3……………….. 133

FIGURE 22. Representativeness of extracted fixed momentary time samples shown as the obtained percentage of time spent in each behaviour \( P_O \) plotted against the overall daily percentage of time spent in that behaviour \( P_D \). ………………………………………………………………………………… 134

FIGURE 23. Representativeness of fixed momentary time samples extracted using the Rule of 1,000 shown as the obtained percentage of time spent in each behaviour \( P_O \) plotted against the overall daily percentage of time spent in that behaviour \( P_D \).………………………………………………………. 138

FIGURE 24. Representativeness of 5-min fixed momentary time samples shown as the obtained percentage of time spent in each behaviour \( P_O \) plotted against the overall daily percentage of time spent in that behaviour \( P_D \)…………. 140

FIGURE 25. Linear regression analysis comparing data measured with graph digitizing software to original data obtained from article authors…………………….. 162

FIGURE 26. Enlarged portion of Figure 25 showing linear regression analysis comparing data measured with graph digitizing software to original data obtained from article authors for data points between 40% and 50% of the session or intervals……………………………………………………………………. 162

FIGURE 27. Cumulative number of data points recorded in bands of 10% (of intervals for discontinuous measures and of the session for continuous measures).………………………………………………………………….. 163

FIGURE 28. Maximum and minimum values reported in each data set (i.e., range)……. 164

FIGURE 29. Mean bout durations from Study 1 (in 10-s bins) and the percentage of behaviours with mean bout durations in each 10-s bin………………. 174

FIGURE 30. Absolute and relative accuracy limits for behaviours occurring for overall durations from 0% to 100% of the time-of-interest………………… 180
FIGURE 31. Example criterion records for 1-hr extracts of low-duration behaviours (1%, 2 %, and 5% overall duration) for varying bout durations............ 187

FIGURE 32. Example criterion records for 1-hr extracts of intermediate-duration behaviours (10% and 20% overall duration) for varying bout durations..... 188

FIGURE 33. Example criterion records for 1-hr extracts of high-duration behaviours (50%, 75%, and 90% overall duration) for varying bout durations........ 189

FIGURE 34. Bubble plot showing samples of 10-s MTS with session lengths ranging from 10 min to 480 min, extracted from 100 simulations of behaviour occurring for 1% of the time in short, medium, and long bout durations.......................................................... 191

FIGURE 35. Bubble plot showing samples of 10-s MTS with session lengths ranging from 10 min to 480 min, extracted from 100 simulations of behaviour occurring for 10% of the time in short, medium, long, and very long bout durations................................................................. 192

FIGURE 36. Bubble plot showing samples of 10-s MTS with session lengths ranging from 10 min to 480 min, extracted from 100 simulations of behaviour occurring for 50% of the time in short, medium, long, and very long bout durations................................................................. 193

FIGURE 37. Bubble plot showing samples of 10-s MTS with session lengths ranging from 10 min to 480 min, extracted from 100 simulations of behaviour occurring for 90% of the time in medium, long, and very long bout durations................................................................. 194

FIGURE 38. Percentage of samples within ± 5% absolute accuracy limits for low-duration behaviours with short bout durations...................... 195

FIGURE 39. Percentage of samples within ± 10% absolute accuracy limits for intermediate- and high-duration behaviours with medium bout durations................................................................................. 196

FIGURE 40. Percentage of samples within ± 5% absolute accuracy limits for low-duration behaviours (1% to 5%)................................................. 197
FIGURE 41. Percentage of samples within ± 10% absolute accuracy limits for intermediate- and high-duration behaviours (10% to 90%)…………………. 198

FIGURE 42. Percentage of samples within ± 5% absolute accuracy limits extracted from simulations of behaviours of 1%, 2%, and 5% overall duration, and short, medium, and long bout durations…………………………….. 200

FIGURE 43. Percentage of samples within ± 10% absolute accuracy limits extracted from simulations of behaviours of 10%, 50%, and 90% overall duration, and short, medium, long, and very long bout durations………………….. 201

FIGURE 44. Mean percentage of samples within ± 10% absolute accuracy limits extracted from simulations of behaviours of overall durations ranging from 10% to 90% across all bout durations…………………………….… 203

FIGURE 45. Percentage of samples within ± 10% absolute accuracy limits extracted from simulations of behaviours of overall durations ranging from 10% to 90% for long and very long bout durations…………………………….. 204

FIGURE 46. Three-dimensional mesh plot from which required session length can be determined when acceptable accuracy limits and practical MTS interval length is selected for behaviours of 10% overall duration………………….. 206

FIGURE 47. Three-dimensional mesh plot from which required session length can be determined when acceptable accuracy limits and practical MTS interval length is selected for behaviours of 20% overall duration………………….. 207

FIGURE 48. Three-dimensional mesh plot from which required session length can be determined when acceptable accuracy limits and practical MTS interval length is selected for behaviours of 50% overall duration………………….. 208

FIGURE 49. Three-dimensional mesh plot from which required session length can be determined when acceptable accuracy limits and practical MTS interval length is selected for behaviours of 75% overall duration………………….. 209
FIGURE 50. Three-dimensional mesh plot from which required session length can be determined when acceptable accuracy limits and practical MTS interval length is selected for behaviours of 90% overall duration……………….. 210

FIGURE 51. Three-dimensional mesh plot from which required session length can be determined when acceptable accuracy limits and practical MTS interval length is selected for behaviours of 50% overall duration……………….. 213

FIGURE 52. Three-dimensional stem plot from which required session length can be determined when acceptable accuracy limits and practical MTS interval length is selected for behaviours of 50% overall duration……………….. 214

FIGURE 53. Range and distribution of bout durations of Study 1 behaviours selected for decision rule testing………………………………………………………….. 217

FIGURE 54. Obtained percentages of time of behaviours with which Study 2 decision rules were tested………………………………………………………….. 219

FIGURE 55. Criterion records for 1-hr extracts of Study 1 data for the behaviours with which the decision rules were tested compared with 1-hr extracts of simulated Study 2 data………………………………………………………….. 220
### LIST OF TABLES

**TABLE 1.** Methods, sub-approaches, alternative names, and methods in the field of work study

**TABLE 2.** Number of observations calculated under Equations 3 and 4 (relative error (RE) and absolute error (AE), respectively) for a range of desired accuracies (1% to 5%) and $P_O$ values

**TABLE 3.** Table of $d$ values (relative spacing) determined under each efficiency value across a range of sampling methods

**TABLE 4.** Reviewed work sampling studies, including details of study length, aim, equation use, sampling method, data collection, and analyses

**TABLE 5.** Operational definitions for each of the adult and participant behaviours for each of the three participants

**TABLE 6.** Behaviours selected and used in additional analyses in Study 1

**TABLE 7.** Behaviours measured for each of the three participants grouped into bins by overall true weekly duration (overall percentage of the week spent engaged in the behaviour)

**TABLE 8.** Percentage of observation sessions in which obtained percentages ($P_O$) fell within the ± 5% accuracy limits of the actual percentage ($P_D$) and the percentage of observation sessions in which behaviour was not detected ($P_O = 0$)

**TABLE 9.** Total percentage of observation sessions in which obtained percentages ($P_O$) fell within the ± 5% accuracy limits of the actual percentage ($P_D$) across behaviours of different weekly duration and when varying $P_O$ values were used
TABLE 10.  Linear regression statistics and Equation 3 error analysis for the Rule of 1,000 and 5-min momentary time sample data........................................ 136

TABLE 11.  Number and percentage of datasets from five volumes of the Journal of Applied Behavior Analysis in which each type of measurement was used................................................................. 160

TABLE 12.  Number of events simulated for each combination of overall duration and bout duration range derived from Equation 7............................ 176

TABLE 13.  Number of observations conducted across an 8-hr day for the range of MTS interval lengths and session lengths................................. 177

TABLE 14.  Upper and lower relative accuracy limits of ± 1% to ± 20% calculated for overall durations ranging from 1% to 90%................................. 181

TABLE 15.  Upper and lower absolute accuracy limits of ± 1% to ±2 0% calculated for overall durations ranging from 1% to 90%......................... 183

TABLE 16.  Overall duration and bout duration values for behaviours from Study 1 selected to be sampled using the decision rules developed in Study 2................................................................. 216
Behaviour analysts are concerned with how people spend their time in particular settings, often with the aim to increase the duration of desirable behaviours and decrease the duration of undesirable behaviours. As such, behaviour analysts must collect data representative of all the time-of-interest in settings to ensure that their clinical decisions are not based on biased samples, and that socially significant behaviour change can be demonstrated. The following thesis sought to identify data-based methods for selecting representative measurement systems for use in behaviour analysis.

Chapter 1 provides a review of the current literature on representativeness in the field of behaviour analysis. Key terms are defined, the factors affecting representativeness detailed, and a discussion of ways to select a sampling method concludes the chapter. Chapter 2 describes work sampling, a method of determining the duration of behaviours in workplace settings used in industrial and organisational psychology. Although now largely of historic interest among organisational psychologists, work sampling methods include the use of equations to determine the number of observations required to obtain a representative sample. The purpose of Study 1 (Chapter 3) was to demonstrate empirically the utility of the work sampling methods for applied behaviour analysis. Study 1 was conducted with naturalistic data were obtained in a school setting. Following Study 1, a comprehensive analysis of a range of values of behavioural duration and sampling parameters was required in order to identify their effect on representativeness. Chapter 4 describes an analysis of the range of reported durations in the applied behaviour analytic literature that was subsequently used to inform Study 2. Study 2 (Chapter 5) simulated recorded streams of behaviour with a range of behavioural durations, from which samples were extracted. The purpose of Study 2 was to describe the effects of some of the dimensions of behaviour and sampling methods on
the representativeness of samples, and subsequently to develop data-based decision rules. The data-based decision rules were developed for use by practitioners to select sampling methods based on estimations of the values of some dimensions of behaviour. The decision rules were tested using the naturalistic data from Study 1. Chapter 5 concludes with recommendations regarding sampling behaviours with a range of durations, and discusses the limitations and implications of the data analysis. The final chapter summarises the findings from Studies 1 and 2, and discusses the importance and challenges of developing a data-based method for selecting representative samples of behavioural duration.
Chapter 1

REPRESENTATIVENESS

Applied behaviour analysis (ABA) involves the application of behavioural technologies derived from research using the science of behaviour analysis to produce change in socially significant behaviours (Baer, Wolf, & Risley, 1968). Naturalistic observations of the behaviours of individuals, in which behaviours are directly observed in the settings in which they occur, are characteristic of applied behaviour analysis (Kazdin, 1979). Therefore, the primary source of data in ABA is direct observation of the interactions of an individual with their environment (Bijou, Peterson, & Ault, 1968). Environmental events and behaviours are defined and measured objectively, without the use of mentalisms or hypothetical constructs to interpret the data (Bijou et al., 1968). It is on the basis of these obtained data that clinical decisions are made and on-going data analysis is used to measure the effects of behavioural interventions and environmental changes. Kazdin (1979) stated that observations of overt behaviour are assumed to be samples of behavioural repertoires, rather than indicative of underlying cognitive processes. There are many factors that may influence how closely a sample of behaviour resembles an overall behavioural repertoire. Thus, practitioners and researchers in behaviour analysis have sought to develop effective methods of observing behaviours in a way to represent accurately a client’s behavioural repertoire.

For accurate inferences to be made about overall behavioural repertoires, observation samples must produce data that closely reflect an individual’s overall behavioural repertoire. In statistics, this is the concept of representativeness, or how closely a sample resembles the whole population (Bertino, 2006). Therefore, the representativeness of the data obtained in
behaviour analytic settings is the degree to which the data obtained from a sample observation session reflects overall dimensions of behaviour, across all opportunities to emit the behaviour across the full time-of-interest (Foster & Cone, 1986). Although texts on behavioural approaches often emphasize the importance of obtaining representative data (e.g., Barlow & Hersen, 1984), there has been little research to show how the dimensions of behaviours (e.g., its duration), and measurement methods affect the representativeness of samples of behaviour. As the task list of competencies issued by the Behavior Analyst Certification Board includes a specific task requiring behaviour analysts to select measurement systems to produce representative data (BACB, 2012), behaviour analysts should have an understanding of the literature on the representativeness of samples of behaviour.

Representative samples are important for data-based decisions in behaviour analysis, such as selecting behavioural goals, selecting interventions, and assessing whether behaviour change has occurred. It may not be possible to observe all occurrences of behaviour and therefore it is important to be able to generalize what is measured in samples to the overall time-of-interest (Brentnall & Bundy, 2009). As samples of behaviour are used to make inferences about overall dimensions of behaviour, behaviour analysis should be concerned about how measurement systems and behaviours affect the representativeness of sampled data. The representativeness of a sample of behaviour is likely to be affected by parameters of the measurement system such as the number, duration, and timing of the start of observation sessions, as well as dimensions of the behaviour such as rate, duration, variability, and temporal distribution (Foster & Cone, 1986).
DEFINITION OF TERMS

ABA is considered to employ measurement techniques more similar to those used in natural sciences than those typically used in social sciences (Wolf, 1978). Therefore, natural sciences can advise on clear definitions of terms pertaining to measurement systems used in ABA. The following terms are important to the concept of data representativeness and will be used as defined below and in the glossary throughout subsequent chapters.

Within the field of metrology, the science of measurement, the definition of ‘measurement’ continues to be refined and reconsidered. Dybkaer (2011) defined measurement as the process by which a number, quantity, value, or qualitative category of what is to be measured (the measurand) is obtained for the purposes of facilitating decision-making. A measurand is analogous to a behaviour or environmental event in ABA. In ABA, one or more dimension of the behaviour will be measured, such as rate, frequency, duration, or inter-response time (Kahng, Ingvarsson, Quigg, Seckinger, & Teichman, 2011). Each dimension of behaviour can be defined by its values; a description of the range of the behaviour (e.g., describing a behaviour as occurring between 15 and 20 times per hour refers to the values of the rate of the behaviour).

Direct observations of behaviour are conducted in order to measure the dimensions of interest. ‘Observation’ refers to a single instance of observing a behaviour (such as the instantaneous observation conducted at the end of each interval in momentary time sampling). The data obtained from a single observation session may be referred to as a ‘sample’ of behaviour.

Producing continuous records of behaviour involves recording a dimension of behaviour such as rate and duration as each instance occurs in real time or from video recordings, whereas discontinuous methods usually involve the division of observation
sessions into intervals of time, with occurrences of behaviour recorded within or at the end of each interval (Kahng et al., 2011). Although some continuous recording may be conducted using pen-and-paper methods such as tallies and stopwatches, continuous recording more often is conducted on electronic devices that record the occurrence, duration, and locus in time of the behaviour. Recording using electronic equipment may facilitate more numerous or detailed analyses of the data (Sanson-Fisher, Poole, Small, & Fleming, 1979). Both types of continuous recording (electronic and pen-and-paper) are of importance in assessing data quality, but distinctions should be made between them when evaluating the literature.

Observation sessions are conducted with consideration of the time-of-interest. Although the time-of-interest may make reference to a particular setting in which a behaviour may occur (e.g., the behaviour of interest may be inappropriate vocalisations only during mathematics lessons), the time-of-interest is better defined as the period of time to be sampled by observation sessions. For example, a school week may be selected as the whole time-of-interest to be sampled by observation sessions if the school timetable is the same week-to-week. Similarly, the whole time-of-interest for nocturnal wandering behaviour may be all sleeping hours across a month in an aged care facility.

Representativeness is assessed by comparing samples to true values, defined as quantitative measures of a dimension of a behaviour that reflect the true state of that dimension (Cooper, Heron, & Heward, 2007). In the field of metrology (measurement science), it has been suggested that, although conceptually logical, obtaining a true value is not often possible in practice (Pavese, 2009). There can also be multiple true values for a measurand (Bich, 2012). García-Santamaria, García-Panyella, and Fuentes-Arderiu (2006) from the field of laboratory medicine suggested that a useful term is ‘conventional true value’ (i.e., what is used as a true value for the purposes of the data analysis).
A conventional true value can be selected in several ways. The conventional true value can be obtained through a reference measure, a value that has been obtained using a measurement system that has been demonstrated to produce true and precise measures of the measurand (International Standards Organization Guide 30, 1992). Another method is to select a value based on convention. For example, the Committee on Data for Science and Technology regularly produces reports on accepted fundamental physical constants for international use (Mohr, Taylor, & Newell, 2008). Conventional true values can also be selected by consensus (e.g., an aggregate or average value from all previous measurements) (García-Santamarina et al., 2006). A conventional true value provides a measure with which to compare the data obtained from direct observation. The error in the data produced from direct observation can be quantified as the difference between the true value and the obtained value (Min & Zhu, 2012), and is often expressed as a percentage or a range. Error can also be conceived as a measure of uncertainty in the measurement (Analytical Methods Committee, 1995). Errors may result from inaccuracies in the measurement system, or in the method of comparing the obtained value to the true value (Pavese, 2009).

Factors affecting the representativeness of a sample can be categorised into those pertaining to the method of measurement, and to the dimensions of the behaviour (Figure 1). Studies evaluating representativeness have focused on comparing different discontinuous methods, determining optimal parameters for discontinuous methods (such as interval duration), and evaluating the effects of different values of the dimensions of behaviour (e.g., varying durations).
Figure 1. Factors affecting the representativeness of samples of behaviour.
Figure 2 illustrates the different possible results in conducting observation sessions of different durations at different times of the day. The first illustrated observation session, 1 hr in duration and conducted at 9 a.m., would produce data that suggest the behaviour is occurring for a large percentage of the day (i.e., is high-duration). However, a shorter observation session, 30 min in duration and conducted at 10.45 a.m., would produce data that indicate a lower-duration behaviour (i.e., occurring for a smaller percentage of the observation session). If the four hours illustrated in Figure 2 are the whole time-of-interest, the behaviour occurs for approximately 50% of this time and neither observation session will produce data that are perfectly representative of the true duration of the behaviour. Therefore, the timing and duration of observation sessions will affect the representativeness of the sample, as will other factors of the measurement system such as the number of observation sessions and the measurement method used (i.e., continuous or discontinuous methods). Similarly, dimensions of the behaviour such as duration and distribution will affect the representativeness of samples.

*Figure 2. Illustration of factors affecting representativeness. Black blocks represent the occurrences (and duration) of the behaviour of interest and dashed lines indicate the timing and duration of observation sessions.*
MEASUREMENT OF REPRESENTATIVENESS

In the few studies that have evaluated representativeness, discontinuous samples of behaviour were compared to continuous records of sessions that were either simulated (e.g., Harrop & Daniels, 1986) or videotaped (e.g., Powell, Martindale, & Kulp, 1975). Powell et al. assumed the continuous record to be a true record of dimensions of behaviour such as rate and duration, and the differences between these values and values obtained from sampling to be the errors in the discontinuous samples. However, although representativeness is most often evaluated by comparing data collected through discontinuous methods to continuous records (true values), continuous records must also be obtained through direct observation and may also contain errors (Johnston & Pennypacker, 2009). There appear to have only been two studies evaluating the representativeness of continuous samples compared with full times-of-interest (Mudford, Beale, & Singh, 1990; Tiger et al., 2013).

Distinctions should be made between the reliability, accuracy, and validity of sampling methods, as these are related but distinct concepts. Accuracy is the degree to which dimensions of behaviour such as rate and duration are reflected in a sample (Foster & Cone, 1986). As true values are rarely measured and used for comparisons in the applied behaviour analysis literature, accuracy is typically determined through comparisons of samples to criterion records obtained through analysis of videotape or similar (Kazdin, 1977). Both measurement systems and observers can be assessed for accuracy. Validity is the degree to which a measurement system measures what it is designed to measure (Johnston & Pennypacker, 2009). Representativeness may be considered a type of validity, because issues in representativeness are issues related to the correspondence of obtained data to true values, or, the generalizability of obtained data to the times-of-interest not sampled.
Behaviour analysis typically uses inter-observer agreement to assess the reliability of the data collected. Reliability is a measure of the stability of the measurement system; the data collected over successive samples will be consistent if behaviour is consistent (Johnston & Pennypacker, 2009). Comparisons of data for the same observation session collected by independent observers measure reliability (Green, McCoy, Burns, & Smith, 1982), but this equates to neither accuracy nor representativeness because high reliability may occur between two inaccurate observers (Lipinski & Nelson, 1974). As Garrett (1942) indicated, comparing the results of one sample to another sample may show agreement, but will not ensure accuracy or detect biases. It is to be expected that successive samples will vary both from overall true values and from each other (Wilson, 1952). Thus, repeated sampling is not necessarily useful for determining representativeness (Garrett, 1942). Representativeness, accuracy, validity, and reliability are all assessed by comparing the data obtained through sampling with data from another source (Kazdin, 1977), however, representativeness is the only measure that requires comparisons to true values. An acceptable measure of behaviour will produce data that are accurate, reliable, valid, and representative.

Representativeness has been discussed relative to the stability of behaviour. Lipinski and Nelson (1974) suggested that decision rules for discontinuing baseline data collection based on the variance in the data collected are helpful to ensure that a representative sample has been obtained. However, this is based on the assumption that data are stable across the whole time-of-interest (e.g., the frequency or duration of behaviour is not variable) and this may not always be the case. Similarly, the stability of data does not directly demonstrate representativeness; stability demonstrates that samples are related but gives no information on whether each sample is representative of the whole time-of-interest.
FACTORS AFFECTING REPRESENTATIVENESS

Consideration of representativeness is necessary when planning data collection for any application in applied behaviour analysis. It is usually impractical to measure behavioural and environmental variables throughout whole times-of-interest. Empirical research may lead to methods for selecting efficient schedules of measurement that produce acceptably representative data.

Each measurement system has advantages and disadvantages that must be considered when selecting which method to use (Haynes & O’Brien, 2000). In an introductory text to scientific research, Wilson (1952) defined systematic errors of measurement systems as errors that are consistent across all observations conducted using that measurement system, and may relate to the variable being measured or be inherent to the method. Identified systematic errors in measurement methods guide practitioners and researchers in choosing a measurement system that will produce data that are sufficiently representative for the purposes of the data collection. Although all measurement systems produce a degree of error (Powell et al., 1975), research has demonstrated that some sampling methods produce systematic errors, or more errors than other methods. What follows is a summary of the factors affecting representativeness, discussed as systematic errors of measurement systems. The factors are unlikely to be mutually exclusive, and may interact to affect representativeness. The numbers of the subheadings correspond to the numbers identifying each factor in Figure 1. All that follows concerns measures of behavioural duration (the overall percentage of time spent engaged in the behaviour across the full time-of-interest), unless otherwise specified.
1. SAMPLING METHOD

1.1. Discontinuous recording

Discontinuous methods of data collection can be categorised into three main methods; momentary time sampling (MTS), partial interval recording (PIR), and whole interval recording (WIR). In MTS, an observation session is divided into intervals at the end of which a decision is made as to whether a behaviour is occurring or not. MTS has been demonstrated to be an alternative to continuous measurement for sampling behavioural duration representatively (Harrop & Daniels, 1986). By contrast, in PIR behaviour is recorded in each interval if it has occurred once or more during that interval. PIR is also known as one-zero sampling (Ary & Suen, 1983). WIR involves scoring an interval if the behaviour occurred throughout the whole interval. In each method, data are often reported as the percentage of intervals: 1. at the end of which behaviour was observed (MTS); 2. during which the behaviour occurred (PIR); or, 3. during which the behaviour occurred throughout the interval (WIR).

In comparing discontinuous methods, Powell, Martindale, Kulp, Martindale, & Bauman (1977) found that MTS produced samples more representative of the durations of behaviours from a continuous record than the samples produced by PIR. Powell et al. suggested that PIR may overestimate duration because a behaviour is recorded as occurring, regardless of how short a duration the behaviour is. If this occurs in many intervals, short-duration behaviours may be recorded in many intervals and the sample may suggest that the behaviour occupies more of the total time-of-interest than it does. In comparison, Powell et al. found that MTS produced both underestimations and overestimations of behavioural duration. Ary and Suen (1983) explained that mixed intervals (intervals in which the behaviour occurs for a proportion of the interval) will affect PIR because such intervals will
always be recorded, but in MTS, sometimes such intervals will be recorded and sometimes they will not, depending on whether the behaviour is occurring at the time the interval ends. Therefore, MTS is as likely to underestimate as to overestimate behaviour (unsystematic error) (Ary & Suen, 1983).

Harrop and Daniels (1986) corroborated the results of Powell et al. (1977) showing that MTS was more likely to produce samples representative of the duration of behaviour than PIR. Meany-Daboul, Roscoe, Bourret, and Ahearn (2007) also found that MTS was superior to PIR in estimating duration, but acknowledged that they had a small dataset. Harrop and Daniels suggested that PIR was more useful than MTS in detecting relative changes in duration, despite producing unsystematic error. Harrop, Daniels, and Foulkes (1990) suggested that although PIR has been recommended when changes duration are to be recorded (e.g., Harrop & Daniels, 1986), MTS may better when estimations of overall duration are required. Suen, Ary, and Covalt (1991) disagreed with the conclusions of Harrop and Daniels, suggesting that MTS is superior because errors are unsystematic, whereas PIR is likely to overestimate duration, particularly when behaviour occur frequently but are low-duration. MTS therefore will remain superior to PIR, even during changes in duration, because error will continue to be unsystematic (Suen et al., 1991). Although this issue has been debated thoroughly by authors such as Suen and Harrop et al., few empirical articles have sought to evaluate these claims and there is a lack of studies that have used obtained data rather than simulated data.

Alvero, Struss, and Rappaport (2008) compared MTS, PIR, and WIR, finding that PIR overestimated duration and WIR underestimated duration of postural safety behaviours. Their results corroborated those of Murphy and Goodall (1980), who found that 10-s WIR underestimated behaviours, particularly stereotypic behaviours of medium duration (bout
duration of between 1.7 s and 13.5 s). In comparison, PIR overestimated behaviours, particularly behaviours of short and medium durations (Murphy & Goodall, 1980). Alvero et al. found that although MTS both under- and overestimated behaviours of different durations, the error in MTS was minimal and that there was little difference in error with differences in duration. However, studies have shown that although MTS produces unsystematic error, the degree of error increases with increased interval duration (e.g., Brulle & Repp, 1984).

1.1.1. Interval duration. Sanson-Fisher, Poole and Dunn (1980) found that the duration of the intervals affects the representativeness of the sample. Powell et al. (1975) found that MTS using interval durations of between 10 s and 120 s produced representative data, but that samples were unrepresentative when intervals of 10 min were used. Similarly, Mansell (1985) found that MTS intervals of up to 30 s produced representative samples, but that 5-min MTS only produced representative samples when the data from 8 hrs of observation sessions were averaged. Alvero, Rappaport, and Taylor (2011) found that increasing MTS intervals to 5 min produced error that was more variable, but that was only marginally larger than the error produced with shorter intervals. However, although the increases in error with increased interval duration were small in the study by Alvero et al., some of the absolute error values across a range of interval durations were large (e.g., 16.17%). The degree of acceptable error is likely to depend on the purpose of the data collection.

Edwards, Kearns, and Tingstrom (1991) also evaluated the duration of MTS intervals (30 s, 5, 10 and 20 min) for behaviours that occurred for 20%, 40%, 60% and 80% of the session. Although Edwards et al. claimed that increasing the interval duration in MTS had little effect on the representativeness of the data, inspection of their reported table of percentage occurrence differences suggests that this may not be the case. The absolute
percentage occurrence difference for a behaviour occurring 60% of the time was up to 18.33% when intervals were 5 min or longer.

Another study that evaluated larger interval durations (e.g., more than 30 s), was Kearns, Edwards, and Tingstrom (1990). Kearns et al. reported that although MTS intervals of $\geq$ 5 min produced more error than shorter intervals of 30 s, there was little difference in the error produced by MTS intervals of 5, 10, and 20 min. These results suggest that if MTS is to be conducted by people in natural settings (e.g., teachers) and longer intervals are unavoidable, longer intervals can be used without increasing error to a large or unacceptable degree. However, both Edwards et al. (1991) and Kearns et al. showed that increasing the MTS interval generally decreased the representativeness of the data.

Brulle and Repp (1984) evaluated the effects of both interval duration and delays in starting each session of MTS. Five behaviours emitted by a child with an intellectual disability were recorded in 30-min sessions across several days. MTS samples were extracted using intervals ranging from 10 s to 40 min. Their results showed that shorter MTS intervals (10 s to 30 s) produced representative samples, whereas increasing the interval duration to longer than 30 s increased both absolute error and error in detecting changes in behaviour. It is important to note, however, that MTS was found to produce representative data when averaged (as MTS is equally as likely to both underestimate and overestimate duration). Brulle and Repp acknowledged that the representativeness of single data points (sessions) may have been poor.

Most studies that have investigated interval duration have kept the intervals a constant duration. Farkas and Tharp (1980) kept MTS intervals a constant duration, but asked some observers to observe multiple participants systematically (i.e., in order) and some observers to observe unsystematically (i.e., to select randomly the next participant to observe). The
results demonstrated that the data obtained through systematic MTS contained less error than data obtained through unsystematic MTS. In unsystematic MTS, although observations were conducted at equal intervals, observations of each participant were conducted at unequal intervals (varying duration based on which participant was selected for observation next). Although Farkas and Tharp presented the data aggregated across participants, the unequal duration of intervals between observations within each participant may have produced more error based on the distribution of behaviours. The behaviours observed were contrived (i.e., the participants were asked to engage in them) were related to test-taking, including adjusting clothing and holding the test paper. As the participants were asked to engage in three behaviours of interest but were not instructed on the duration, timing, or frequency of each behaviour, it is unknown how the behaviours were distributed. As observations were conducted at random intervals, the distribution of the behaviours may not be expected to affect the representativeness of the sample. However, observers may have unintentionally selected which participant to observe based on how salient their behaviour was (e.g., when a participant changed behaviour or was obviously adjusting their clothing). To avoid this potential confound, the observers could have been given a randomised schedule of who to observe when. Therefore, constant or randomised durations of intervals may affect the representativeness of the data obtained, particularly in relation to the distribution of behaviour. For example, constant interval observations of cyclic behaviours may result in overestimations of behaviour if observations are conducted at the same point in the cycle.

Ary and Suen (1983) suggested that the optimal interval duration will be shorter than the shortest duration of the behaviour and the shortest time between occurrences of the behaviour (inter-response time). They suggested that information on the average duration and inter-response times should be gathered before recording the behaviour and should be
used to select interval duration. Sanson-Fisher et al. (1980) also recommended conducting initial observations through continuous recording in order to determine appropriate interval duration. Despite suggestions to use data to determine interval duration, few studies report methods to do so. Although there does not appear to have been a more recent evaluation of commonly-used interval durations, Kelly (1977) found that 10-s intervals were the most commonly reported in *Journal of Applied Behavior Analysis*. Although 10 s is a short interval duration and has been shown to produce representative data (e.g., Brulle & Repp, 1984), tailoring aspects of a measurement system (such as interval duration) to the purposes of data collection is more desirable than arbitrarily selecting the measurement method.

1.2. Continuous recording

Continuous methods of recording behaviours, in which all instances of the behaviour are recorded during an observation session, can provide a more comprehensive measurement than discontinuous methods such as interval recording (Cummings & Carr, 2009). However, relatively fewer metrological studies have been conducted on continuous methods, despite these methods being used in many research articles (Mudford, Taylor, & Martin, 2009). As it is not always possible to record the whole time-of-interest of a behaviour, continuous methods can be used to conduct observation sessions which produce samples of the whole time-of-interest. Therefore, although the issue of representativeness applies to both continuous and discontinuous methods of sampling, there appears to be very little research evaluating the representativeness of continuous samples by comparing them to full times-of-interest.

Mudford, Beale, and Singh (1990) and Tiger et al. (2013) have conducted studies evaluating the representativeness of continuous samples. Mudford et al. measured six mutually-exclusive behaviours of adults with intellectual disabilities, including social
interactions and inappropriate behaviours. Behaviours were recorded on a computer in 150-min blocks for each participant, after which the computer was used to extract samples ranging from 15 to 135 min (in increasing 15-min blocks) from the whole time-of-interest (150-min session). The starting point of the simulated samples was varied between the beginning of the session, the midpoint of the session, ending at the end of the session, and a randomly-selected start time. When they compared samples to the whole time-of-interest data, they found that shorter samples produced more error in recording duration than longer samples, particularly for lower-duration behaviours (i.e., those occupying 1% to 3% of the session). They also found little difference in the representativeness of samples started at different times during the session.

Tiger et al. (2013) recorded children’s behaviour through four whole school days. From continuous, adjacent 10-min records of behaviour, Tiger et al. extracted 10-, 20-, 30-, and 60-min samples. The percentage of time spent engaging in problem behaviours measured in the extracted samples was compared to the true daily percentage of time spent engaged in the behaviour. They found that behaviours with low variability were sampled representatively in most 10-min continuous samples, but high-variability behaviours were not. They also found that increasing the sample size increased representativeness (although only marginally), corroborating the results of Mudford et al. (1990).

1.3. Observation session duration

Mudford et al. (1990) suggested that observation session duration should be determined empirically, because it has been demonstrated to affect the representativeness of data. Since, there has been little research conducted that guides practitioners and researchers on empirical methods to determine the optimal observation session duration.
Although some studies have conducted relatively longer observation sessions (e.g., 10 hrs; Mansell, Jenkins, Felce, & de Kock, 1984), most studies have conducted observation sessions of much shorter duration. For example, Powell et al. (1977) used 30-min sessions and Alvero et al. (2008) used 15-min sessions. As an example of use of relatively longer observation sessions, Landesman-Dwyer and Sackett (1978) conducted 24-hr observation sessions of sleep and levels of activity in children with profound intellectual disabilities. Observations were conducted every 10 min, providing detailed data on behavioural patterns.

Fewer studies still have evaluated the effect on representativeness of changing the observation session duration. Leyendecker, Lamb, and Schölmerich (1997) found that four or more consecutive 45-min blocks of time sampling were required for the observation sessions to be inter-correlated for one category of directly observed mother-infant behaviours. Leyendecker et al. suggested that longer observation sessions were therefore required for data to be stable, and thus representative. However, comparing observation sessions to other observation sessions is not a true measurement of representativeness and no comparisons are made with true values. As Leyendecker et al. used discontinuous sampling (i.e., observers observed for 20 s and recorded for 10 s), the number of observations increased with increases in observation session duration. They did not evaluate optimal combinations of observation session duration and interval duration. It may be that more frequent observations conducted across shorter observation sessions produce more representative data than widely-spaced observations across longer observation sessions.

Devine, Rapp, Testa, Henrickson, and Schnerch (2011) simulated behaviours of durations ranging from 25% to 75% of the session and simulated small, medium, or large changes in duration across settings. They evaluated the effect of varying observation session duration and interval duration, finding that in general, more representative data were obtained
with longer observation sessions. They also found that MTS was superior to PIR in detecting changes in behaviour, and that increasing the MTS interval duration to 30 s produced representative data if the session duration was increased to 30 min or 60 min. Shorter observation sessions (10 min) produced representative data if shorter MTS intervals (10 s) were used.

Mansell (1985) found that when the time between MTS intervals was increased to 5 min, 8 hrs of observation were required for representative data to be obtained for behaviours occurring for 25% or more of the session. High-duration behaviours (i.e., behaviours occurring for 45% or more of the session) were more representatively sampled than low-duration behaviours. Shorter observation sessions with more frequent sampling (MTS 30 s) produced similar results. It has been suggested that simply increasing the sample size of questionnaire respondents may not increase the utility of the information (Blalock & Dial, 1990). The same may be true of direct observation data; longer observation sessions may not yield more representative data if factors such as interval duration or the dimensions of the behaviour have not been considered.

Mansell and Beadle-Brown (2011) found that conducting MTS observations every 20 s for 2 hrs produced a sample representative of how residents of a residential institution spent their time across an 11-hr day. However, the 2-hr observation session produced over-estimations of two behavioural categories. As the 2-hr observation session was conducted over a period that included a mealtime (a time during which one-on-one interaction, the category overestimated, would be expected to increase), this illustrated the effect that start time of observation sessions can have on the representativeness of samples, regardless of the duration of the observation session. Le, Perlman, Zellman, and Hamilton (2006) found that although a single 2-hr observation session predicted child-staff ratios measured over ten 8-hr
observation sessions, the time of day affected the representativeness of the sample (i.e., the child-staff ratio decreased in the afternoon). Therefore, both the timing of the start and the duration of observation sessions must be considered in choosing an observation schedule.

In contrast, Brentnall, Bundy, and Kay (2008) reported that the timing of the start of observation sessions of child playfulness did not affect the scores on an observational rating scale. They also found that increasing observation sessions from 15 min to 30 min did not increase the representativeness of the sample. Although these results show little differences in representativeness of either the timing of the start of or duration of observation sessions, Brentnall et al. used a specific observation tool and the timing of the start of observation sessions was only varied by 15 min. In addition, representativeness was not directly evaluated; each sample was compared to the other samples to test for correlations.

Although generally it is suggested that longer observation sessions will produce more representative data (Johnston & Pennypacker, 2009), conducting lengthy observations can be difficult in applied settings if it is not cost-effective or observers are not available (Mansell et al., 1984). Even when lengthy observation sessions are possible, factors such as observer fatigue and the possible resulting decrease in the quality of the data should be considered. Smith, Madsen, and Cipani (1981) found that observers produced less reliable data with longer observation sessions (e.g., 2 hrs), but did not explore or report an optimal observation session duration to avoid unreliable data collection.

1.4. Timing of the start of observation sessions

Cooper et al. (2007) suggested that poorly scheduled observation sessions may affect representativeness. For example, factors such as time of day, the presence or absence of others, location, and the activity within which behaviour is recorded are all variables that may affect the representativeness of the sample. For example, data obtained during a meal at a
restaurant may not be representative of rates of problem behaviour that typically occur during the evening family meal in the home. The issue of when to schedule representative observation sessions includes both where to observe the behaviour (the setting), and the timing of observation sessions across a day or week.

Data collected in analogue or laboratory settings may not be representative of behaviours that occur in natural settings. For example, Rhule, McMahon, & Vando (2009) found parents rated some tasks used to measure parent-child interactions in the laboratory as producing interactions representative of home interactions, but some tasks as producing less representative interactions. Similarly, Pett, Wampold, Vaughan-Cole, and East (1992) showed that mother-child interactions were inconsistent across two different settings and thus the data collected in each setting produced different results.

Specifying the setting in which the behaviour of interest occurs is important (Bijou et al., 1968), and will affect decisions regarding measurement and data collection. Bijou et al. also identified behavioural dimensions, the purpose of data collection, and practical constraints as factors affecting the decision of when to schedule observation sessions. Observation sessions conducted across a range of settings may produce unstable or inconsistent data (Doll & Elliott, 1994), but if the purpose of the data collection is to evaluate the behaviour across a range of settings, inconsistency of data across settings is a finding rather than a methodological issue. Data to show that behaviour occurs in some settings and not others are useful to behaviour analysis. The purpose of data collection will greatly influence the choice of settings in which behaviour is to be observed.

In natural settings, the timing of observation sessions can affect the representativeness of the data. Russell and Bernal (1977) found that the rate of children’s challenging behaviour varied across a week, decreasing as bedtime approached and increasing on cold days. In
contrast, Le et al. (2006) found that the day on which the child-staff ratio in an early childhood setting was measured had no effect on the representativeness of the data on child-to-staff ratios, but that the time of day did affect representativeness. As there was little difference in child-staff ratios between days, but Le et al. found differences in the ratios within days (more staff per child in the morning), the distribution of the ratios was uniform across days. Therefore, if the aim of data collection was to evaluate how the child-staff ratios changed across a day, samples were best scheduled across the whole day. If the aim of the data collection was to determine the average child-staff ratio, observation sessions conducted at the same time each day are unlikely to provide a representative sample. Although Le et al. measured child-staff ratios, not behaviour, their results show how the timing of observation sessions can affect representativeness.

To address the issue of when to schedule observation sessions, Russell and Bernal (1977) recommended that all observation sessions are conducted at the same time of day or week in order to be confident that changes in the behaviour are due to the intervention implemented. If this is not possible, however, observation session time should be randomised so that error is unsystematic.

1.5. Number of observation sessions

Although Johnston and Pennypacker (2009) suggested that observation sessions should be as long as possible and repeated as many times as is possible, this is unlikely to be practical for many practitioners. Therefore, studies that show how many observation sessions are required to obtain a representative sample for behaviours with particular dimensions are important. For example, Osterhaus and Passchier (1992) determined that at least three weeks of self-recording of the frequency and duration of headaches was required to account for 90% of the variance in the data. Doll and Elliott (1994) measured categories of pre-schoolers’
social behaviours across nine observation sessions, using the summed data from all nine
observation sessions as a whole time-of-interest with which to compare various numbers of
observation sessions. They found that at least five 10-min observation sessions were required
to observe proportions of time spent in each category that were representative of the data
from all observation sessions. However, the nine observation sessions comprising the whole
time-of-interest with which different numbers of observation sessions were compared were
conducted across multiple days. Data were obtained with discontinuous methods, therefore
representativeness of true values was not evaluated (i.e., sample data were not compared to a
full time-of-interest measure of behaviour).

Similar methods were used by McKeivitt and Elliott (2005), who demonstrated that
three 30-min observation sessions produced data representative of categories of children's
social behaviour. They found that the data obtained from early observation sessions could be
used to predict frequencies of behaviour obtained in later observation sessions. However,
they found that more than three observation sessions did not further increase the accuracy of
prediction. McKeivitt and Elliott did not obtain true values for comparison and therefore their
data could only be used to predict the data obtained from future observation sessions but
could not assess representativeness. Comparing samples with other samples may inform the
observer about the stability of behaviour or about the test-re-test reliability of measurement
systems, but if the samples are not representative, comparing samples to samples is of little
use. Moore (1998) found that between sixteen and twenty 8-min observation sessions
produced representative data, but that conducting between 20 and 30 observation sessions did
not increase representativeness. However, Moore conducted all 8-min observation sessions
consecutively using a combined MTS / PIR procedure, and used the means and standard
deviations of behaviour as criterion measures. Therefore, no true values were obtained through continuous recording; samples were compared to a longer discontinuous sample.

In summary, few studies have evaluated the number of observation sessions required to obtain a representative sample. There has been little research exploring the interaction between observation session duration and the number of observation sessions (i.e., whether more numerous, shorter observation sessions produce more representative data than fewer, longer observation sessions). The purposes of the data collection will likely affect the number of observations required and able to be conducted (e.g., a practitioner may only be able to obtain a few baseline observation sessions before an intervention is to be implemented).

1.6. Observer accuracy

Repp, Nieminen, Olinger, and Brusca (1998) identified seven major factors that can affect the accuracy of the data produced by observers: 1. reactivity; 2. observer drift; 3. feedback; 4. sampling method; 5. sampling location; 6. participant characteristics; and, 7. reliability. The effect of each of the factors has been demonstrated in empirical studies. For example, reactivity, during which the presence of an observer changes the behaviour of the person being observed, was demonstrated to occur differentially across participants by Schonwetter, Miltenberger, and Oliver (2014). They found that the presence of observers increased the number of laps completed by some swimmers. Observer drift, in which an observer gradually changes their application of the behavioural definition to what is being observed, has been demonstrated to occur across pairs of observers (e.g., Kent, O’Leary, Diament, & Dietz, 1974).

Data may be collected via scoring video tapes or in-situ observations. There is little research comparing the representativeness of data collected by the two methods, however an
advantage to video recording is the ability to re-watch the video to check for accuracy and have multiple observers record data from the same video. Mathiassen, Liv, and Wahlström (2013) showed through a cost analysis that video recording could be more time efficient and produce more data than in-situ observations as multiple observers could record data from the same video. However, agreement between observers does not necessarily equate to accurate data. In addition, some behaviours may be difficult to capture on video, such as fine motor movements or auditory-based behaviours (Repp et al., 1998). Therefore, sampling location, in addition to the effects of sampling method (e.g., MTS compared with PIR) as previously described can affect observer accuracy.

Providing feedback to observers on the accuracy of their recording has been shown to bias subsequent recording. For example, Lerman et al. (2012) found that providing verbal feedback and monetary rewards could bias recording towards missing the occurrence of behaviour or falsely recording the occurrence of behaviour, depending on the feedback and contingency for reward. Similarly, they found that behaviours identified as ambiguous (compared to the operational definition of the behaviour) were less likely to be recorded accurately by observers. The ambiguity of behaviour was classified as a participant characteristic by Repp et al. (1998), along with the predictability and pattern of responding.

The reliability of data collected by direct observation in applied behaviour analysis is most frequently evaluated by calculating inter-observer agreement (IOA). However, observers have been shown to be more accurate in recording data when other observers are present than when they are alone (Fradenberg, Harrison, & Baer, 1995), and the degree of inter-observer agreement can also be affected by the algorithm used to calculate IOA. For example, Mudford, Martin, Hui, and Taylor (2009) found that some algorithms were likely to
inflate IOA, whereas others were overly stringent for behaviours of particular duration or rate.

As observer accuracy can be affected in the aforementioned ways, the resultant inaccurate data can be unrepresentative of true values. For example, if the presence of an observer provides an additional discriminative stimulus for attention and increases the duration of attention-maintained behaviour, the duration of the behaviour may be overestimated. The sample will therefore be unrepresentative of the time-of-interest (during which the observer is not usually present). Similarly, if an observer is less accurate in recording the duration of behaviour as they are observing alone, the resultant data may under or overestimate true duration, and are therefore not representative of the time-of-interest. Factors affecting the accuracy of observers can be addressed by ensuring thorough and adequate observer training, ensuring that observation is as unobtrusive as possible, allowing for an adjustment period for those being observed, and recording behaviour from permanent products such as videos (Repp et al., 1998).

2. DIMENSIONS OF BEHAVIOUR

Dimensions of behaviour such as rate, duration, variability, bout duration, and inter-response time are likely to affect representativeness. It has been demonstrated that data obtained on the rate and duration of the same behaviour may not provide the same information (Saunders, Timler, Cullinan, Pilkey, Questad, & Saunders, 2003). As behaviours have multiple dimensions, it may be useful to choose the dimension of interest (e.g., what is to be increased or decreased) and choose a sampling method that has been shown to be useful in obtaining data that represent the dimension of behaviour.
Duration. Gardenier, MacDonald, and Green (2004) selected duration as the dimension of interest for stereotypic behaviours. They found that MTS produced more representative samples of low, moderate, and high durations of stereotypy than PIR (when sampled from continuous records of the behaviour). Specifically, MTS both under- and overestimated all durations (although mostly underestimated), and PIR consistently overestimated durations. The representativeness of the samples was affected by both the duration of the behaviour and the size of the MTS intervals. MTS with 30-s intervals produced the most error for stereotypy of low and moderate durations (behaviours occupying ≤ 40% of the session). Shorter MTS intervals of 10-s and 20-s intervals produced smaller errors. Although PIR produced more overall error than MTS, PIR produced more representative samples of low-duration stereotypy than of moderate- or high-duration stereotypy.

Gardenier et al. (2004) corroborated prior research that showed increasing interval duration increased error in recording (e.g., Powell et al., 1975), and demonstrated that duration affected the representativeness of sampling. Saudargas and Zanolli (1990) also found that duration affected the representativeness of samples obtained with MTS. Although they found that MTS with 15-s intervals produced representative samples of behaviours of a range of durations in a naturalistic setting (e.g., out-of-seat behaviour), they also found that short-duration, low-rate behaviours were less likely to be sampled representatively. Similarly, Harrop and Daniels (1985) suggested that short-duration behaviours will not be sampled representatively unless the behaviour occurs at a high rate, and that increasing MTS intervals to 30 s or more is unadvisable if bouts of behaviour are short in duration (e.g., 20 s). MTS has been suggested to be useful when bout duration is long, even when inter-response time is short and bouts are infrequent (Harrop, Murphy, & Shelton, 1994).
Mudford et al. (1990) also varied the duration of the behaviours sampled. When they varied observation session duration relative to total time-of-interest, high-duration behaviours (i.e., occupying \( \geq 50\% \) of the session) were more representatively recorded than low-duration behaviours as interval duration decreased. Similarly, Powell et al. (1977) found less measurement error for samples of high-duration behaviours, with the error further decreasing with decreased interval durations.

Rate. Repp, Roberts, Slack, Repp and Berkler (1976) simulated low-, moderate- and high-rate behaviours (rates of 0.1, 1, and 10 behaviours per minute, respectively), and found that time sampling slightly overestimated low-rate behaviour but produced samples that were poorly representative of moderate- and high-rate behaviours (mostly underestimated behaviour). Repp et al. also simulated behaviour as occurring at a constant rate throughout the session and behaviour that occurred in bursts (e.g., instances of behaviour were unevenly distributed across the session). Time sampling was found to overestimate unevenly distributed behaviour more than behaviour that was constant, particularly when behaviour was high-rate. However, rather conducting brief MTS observations end of each interval, Repp et al. recorded behaviour for 5 s at the end of each interval. Therefore, the sampling method used was not a true MTS and was not a method typically reported in the literature. The effects of rate on the representativeness of samples taken through MTS, may therefore be different to those reported by Repp et al (1976).

Thomson and Baer (1974) evaluated the representativeness of samples of a low- to moderate-rate behaviour and a moderate- to high-rate behaviour. For two out of three participants, samples of moderate- to high-rate behaviour contained more error than samples of low- to moderate-rate behaviour. However, the opposite was true for the third participant. The sampling methods used by Thomson and Baer, however, involved combination of time
sampling in continuous blocks and was therefore dissimilar to standard discontinuous
sampling procedures such as PIR and MTS. Very low-rate behaviours may not be suitable
for discontinuous sampling at all; perhaps it is better to use continuous recording of
occurrences for such behaviours (Adams, 1991). Studies such as Repp et al. (1976) and
Thomson and Baer (1974) have demonstrated that dimensions of behaviour such as rate and
duration influence the representativeness of samples, and the results of Repp et al. suggest
that the distribution of behaviour is also an important factor.

**Distribution of behaviour.** Algase, Kupferschmid, Beel-Bates, and Beattie (1997)
conducted a study to assess the effects of the distribution of behaviour on sampling. Algase et
al. showed that wandering behaviour in patients with dementia was random in both bout
duration and number of instances. Therefore, a 2-hr sample of this behaviour was sufficiently
representative of wandering for a 24-hour period. Studies using simulated data such as
Kearns et al. (1990) typically have generated occurrences of data that are randomly
distributed across the time-of-interest. Although this is useful to assess other factors affecting
representativeness such as interval duration and the duration of behaviour, no studies have
compared the utility of sampling methods for behaviours that are distributed uniformly (e.g.,
occur at the same times each day) or cyclical (e.g., occur in a repeating pattern or order). For
example, stereotypy that always occurs at a higher-duration in the morning and decreases in
duration across the day may be distributed uniformly, and assembly line work in a factory
may be cyclic in nature.

**Changes in behaviour.** Just as rate, duration, and distribution of behaviours are
important to consider in choosing a representative measurement system, it is also important to
select a measurement system that will continue to produce representative samples of
behaviour when behaviour change occurs (e.g., during an intervention). The ability of a
sampling method to detect changes in behaviour can be referred to as sensitivity (Harrop & Daniels, 1986).

Kearns et al. (1990) simulated data of decreasing duration across four sessions. They showed that when behaviour decreased from occurring for 80% of a session to 40% of a session, MTS with intervals of varying durations was able to detect changes in duration. Longer intervals (e.g., 5 to 20 min) were able to reflect the trend of the behaviour change, but produced less representative samples than shorter MTS intervals of 30 s. MTS with longer intervals appeared to produce more representative samples of increasing trends than decreasing trends. However, although they showed MTS could produce representative samples of behaviour changing in duration, the behaviour change was consistent across sessions (e.g., the behaviour increased or decreased in duration by exactly 20% each session). Behaviour change in natural settings is unlikely to be as systematic or consistent, and therefore the utility of sampling methods for producing data representative of more realistic behaviour change should be evaluated.

To evaluate further the utility of MTS and PIR with different interval sizes to detect changes in rate and duration, Rapp, Colby-Dirksen, Michalski, Carroll, and Lindenberg (2008) used simulated data to compare MTS and PIR samples with continuous records of behaviour. Rapp et al. found that MTS with intervals of 10 s, 20 s and 30 s produced representative samples of moderate and large changes in duration, but that PIR with 10 s or larger intervals did not in all instances. Alternatively, 10-s PIR was found to be more accurate than MTS at detecting changes in rate, although MTS became more accurate in detecting changes at higher rates. Rapp et al. concluded that for behaviours occurring more than 20% of the time, MTS with small intervals was the most representative measure and that continuous recording was best for measuring small changes.
Devine et al. (2011) evaluated increasing MTS interval duration as observation session duration increased, as they hypothesised that larger intervals would be sufficient to detect changes in longer observation sessions. Their results supported their prediction; 80% of changes in duration during 30-min and 60-min observation sessions were able to be identified with 30-s MTS. Devine et al. corroborated reports that PIR and MTS are more suitable for detecting changes in frequency and duration, respectively (Rapp et al., 2008). The results obtained by Devine et al. support the use of longer MTS intervals with longer observation sessions, however 30-s MTS is labour-intensive and may be impractical in some settings.

SELECTING SAMPLING METHODS

Choosing a sampling method will depend on a number of factors such as the practical constraints (e.g., availability and skill of observers), the representativeness of the data produced by the method, and the purpose of data collection. From the field of ergonomics, in which work productivity is increased through measurement of work behaviours, Rezagholi and Mathiassen (2010) suggested that measurement systems should be selected based on the appropriateness of the method for the purposes of the study. The issue of efficiency and cost-effectiveness of measurement systems relative to the quality of the data produced has also been considered in both ergonomics and behaviour analysis. For example, Rezagholi and Mathiassen stated that cost-efficient systems may not yield representative accurate data, just as obtaining data with small errors may not be cost-efficient. Therefore, the practicality and effort required in using a measurement system must be considered.

Although research comparing sampling methods will guide practitioners on choosing a method that will produce representative data, it is recommended that details such as the
number, duration, and timing of observation sessions required for representative data be selected following analyses in the observation setting (Alevizos, DeRisi, Liberman, Eckman, & Callahan, 1978). Rapp et al. (2008) provided guidelines for practitioners to choose both the sampling method and interval size. It was suggested that practitioners should conduct preliminary observation sessions of approximately 10 min to determine the values of the behavioural dimension to be increased or decreased (e.g., rate or frequency) and use the information to choose a method (Rapp et al., 2008). Although this information is useful, amalgamating the results of this study and others into a flowchart may be of particular use to practitioners for quick, research-guided reference. Fiske and Delmolino (2012) provided such a flowchart for practitioners to select discontinuous sampling methods based on the desired behaviour change. Further research into the development of practical tools to select sampling methods would be of benefit.

MTS has been suggested to be preferred by observers and is likely to produce smaller errors in recording than PIR, regardless of the complexity of the measurement system (e.g., number of behaviours to be recorded) (Murphy & Harrop, 1994). As MTS has been demonstrated to produce representative samples of behavioural duration and PIR with small intervals has been shown to produce representative samples of frequency (Rapp et al., 2008), the dimension of the behaviour of interest will guide the choice of sampling method. If the behaviours of interest occur at low rates, time sampling methods that produce representative samples of high-duration or frequency behaviours may not be appropriate (Alevizos et al., 1978). Measurement methods that are able to record the contingencies of low-rate or duration behaviours may be of more use.

Although minimising error is desirable, marginal increases in error with increasing interval duration in MTS may be acceptable if increased interval durations increase the time
efficiency of data collection. For example, increased sample duration may increase the number of participants that can be observed or permit the observer to engage in other behaviours between observations (Alvero et al., 2011). For data collection to be conducted in natural settings by teachers, parents and other people in the setting, MTS with long intervals that is easy to conduct and will produce less observer error may be most appropriate. Similarly, MTS conducted at long intervals over a full day may produce data that is more representative than MTS with short intervals conducted over only a portion of the day. Alvero et al. also suggested that if the sampling method is easy to use and efficient, it may be used more frequently. Increasing the social validity of a measurement system may increase the amount and quality of the data collected by those using it. Observers using MTS will spend less time observing than when using other methods (Kearns et al., 1990). Therefore, methods such as MTS permit multiple behaviours or participants to be observed within an observation session and may be used by non-professionals (Alevizos et al., 1978). Care must be taken, however, in observing multiple people or behaviours, because increasing the complexity of measurement systems has been demonstrated to reduce the reliability of observational data (e.g., Jones, Reid, & Patterson, 1974). Although Jones et al. evaluated reliability, which is not a direct measure of representativeness, representativeness is likely to be affected if one or more observers is not recording accurately. Issues of observer error will affect the representativeness of samples of behaviour (as identified in Figure 1).

SUMMARY

Overall, representativeness is affected by a number of factors including sampling method, the timing of the start of observation sessions, and the number of observation sessions. Although there is some literature to guide practitioners in choosing a measurement
system, the representativeness of the data will largely depend on the time-of-interest and the purposes of the data. Similarly, the level of acceptable representativeness will depend on the purpose of data collection (Fassnacht, 1982; Moore, 1998). Representativeness is not commonly operationally defined (Bar-Hillel, 1980) and therefore can be subjective. Although Bar-Hillel suggested that a quantitative measure of representativeness should be developed, the degree of desired representativeness will vary between uses. Similarly, few studies or practitioners measure representativeness prior to or during a study to ensure that the data represent true values; often reliability measures are the only test of data quality. Representativeness should be considered for all data collection, as data that are valid, reliable, and accurate are unlikely to be of much use if not representative of the time to which generalisations are to be made. This may be best done by evaluating a small sample to determine the dimensions of the behaviour and using empirical research to guide the selection of a measurement system (e.g., Rapp et al., 2008).

In summary, a review of the literature generates some general rules regarding factors that increase the representativeness of a sample: 1. choosing MTS over PIR and WIR; 2. shorter intervals in discontinuous recording; 3. longer observation sessions; 4. more observation sessions; 5. careful consideration of the timing of the start of observation sessions; 6. higher-duration behaviours are sampled more representatively with both continuous and discontinuous recording; 7. and, sampling methods should be selected empirically.
Chapter 2

WORK SAMPLING

REVIEW OF WORK SAMPLING METHODS

Hypothetical Case Example

The following hypothetical case is derived from published studies in the field of nursing. The management team of a large public hospital were concerned that the average time it took for patients to be processed in the emergency department was long. In addition, the hospital had received numerous complaints and bad publicity regarding the efficiency of the emergency department. In an effort to reduce processing times and improve the efficiency of the department (i.e., the speed at which patients were processed whilst a high standard of care was maintained), a work sampling study was ordered. A work sampling expert was contracted to conduct a study to show how nurses in the emergency department were spending their time. Work sampling was selected as a time-efficient way of gathering information through objective observations rather than subjective self-reports. Emergency nurses were selected for the study because they were the first point of contact for patients in this hospital and were responsible for triage. Therefore, how nurses spent their time was thought to be influencing the processing times of patients.

The work sampling expert met with the management team to discuss how the study was to be carried out. Management informed the expert that in order to minimise disruption to patients and staff and to minimise the cost of the study, the expert was to conduct the study over one month. One month was also selected so that the time nurses spent compiling monthly summary reports could be included. The month chosen for the study was March because this was neither a busy nor quiet period for the emergency department (based on data...
provided by the management team). The work sampling expert, in collaboration with hospital staff, created and defined a list of behavioural categories to be measured.

The work sampling expert conducted a preliminary study for one day, which provided an estimate of how nurses were spending their time. These estimates were then entered into a work sampling equation to determine how many observations were required in the study for a representative sample. An observation was a momentary evaluation of which of the behavioural categories in which the nurse was engaged. At the end of a day of study, the proportion of observations in which each behaviour was observed was calculated. Each day during the study, the proportions of time spent in each behavioural category were calculated and plotted on a control chart; a graph that allowed the expert to see if the proportions varied greatly from day to day. The number of required observations was also recalculated daily using the new obtained proportion values.

At the conclusion of the month of observation, overall proportions of time spent in each behavioural category were calculated and presented to the management team. The management team concluded that the nurses were spending more time doing paperwork and computer work than was desirable, and less time in direct patient care. Hospital management introduced portable computer tablets on which the nurses could record information (to reduce the time required to walk to a computer and complete paperwork), as well as reviewing and combining several forms to reduce the amount of paperwork required from nurses. The nurses reported increased available time to spend with patients, and management contacted the work sampling expert to repeat the study in order to measure changes in the proportions of time spent in each category. The current hypothetical case study will be used to illustrate each component of work sampling throughout the following review.
BACKGROUND

Industrial engineering, an approach which integrates mathematical and scientific principles in the analysis, design and adjustment of industry operations, is derived from Frederick Winslow Taylor’s ideas of scientific management (Nadler, 1955). Taylor (1911) suggested that inefficiencies in industry could be systematically analysed, and that the management of such inefficiencies could be applied scientifically to a range of activities. Taylor (1911) argued that a goal of management is to maximise the efficiency of workers, and that management should be able to give workers clear instructions and clear expectations of the amount and type of work they are to complete in a specified period of time. The actions of workers can be broken into components, analysed and modified, leading to more productive work. Therefore, the systematic analysis of the most efficient and effective way to assess and manage a process could be argued to be a fundamental goal of industrial engineering. Further pioneering work in the 1910s by Frank and Lillian Gilbreth (an engineer and psychologist respectively) analysed the methods used in industry, including the classification and study of specific human body motions they labelled ‘therbligs’ (Karger & Bayha, 1965). Time study, in which the standard time taken to complete an action is measured and motion study, which seeks the most effective way to complete a task by analysing the movements, were therefore initially distinct. Karger and Bayha (1965) suggested that although followers of Taylor’s time study and the Gilbreths’ motion study groups initially disagreed with each other’s method, the field eventually combined to form time-and-motion study. Despite the apparent historical disagreements between the two approaches, it has been suggested that time study and motion study complement each other and are underpinned by the same goal, to evaluate the most effective way of completing a certain task (Morrow, 1946).
Time-and-motion study is a branch of industrial engineering that is underpinned by four general aims; to reduce work to only what is necessary, to establish the most efficient order in which to complete the work, to develop standards for the work, and to develop time standards for the work (Nadler, 1955). It also provides the basis for training operators to perform the analysed task in the most efficient way (Barnes, 1964). In reading the texts written in the 1950s and 1960s, a period in which there was an academic focus and the Operations Research Society of America and The Institute of Management Sciences were founded (Saunders, 1982), it becomes apparent that the field of time-and-motion study uses a number of slightly differing (and hence confusing) terms. Therefore, Table 1 collates information from a number of books written in this period (Barnes, 1964; McCormick, 1982; Morrow, 1946; Morris, 1969; Niebel, 1967).

‘Work study’ or ‘operation analysis’ is an umbrella term used to describe the field of time-and-motion study. Within work study are three areas, method study (which includes motion study), work measurement (which includes time study), and job evaluation (Morrow, 1946). Although all of the sub-approaches of work study are widely applicable and complementary, work sampling, a method of determining how workers spend their time, has been adopted as the underlying approach in empirical studies of how time is spent across a range of areas. The utility of a scientific, efficient, and time-economical way to determine how time is spent across a range of activities could be of interest to any area in which the durations of activities are important. Work sampling, initially referred to as activity sampling, was first developed by Tippett (1935) as a quantitative method of producing cost-effective but accurate information on how people were spending their time in a textile factory. Based on statistical sampling techniques and on the laws of probability, work sampling assumes that taking ‘snapshot’ observations of the activities people are engaged in will be indicative of overall durations of those activities. The subject of the study is observed
at random or fixed times throughout the day and their work behaviour recorded from a list of pre-determined categories (Heiland & Richardson, 1957). This permits the person recording the subject’s behaviour to make inferences regarding overall durations of behaviours without having to record data for the full time-of-interest (for example, a work day or week). Niebel (1967) identified that, in addition providing a time-effective way of measuring work behaviours, work sampling has the additional benefits of reducing discomfort in the worker that may result from long periods of being observed. Work sampling also allows one observer to observe multiple workers in one session.

The process of designing and conducting a work sampling study has been detailed in many introductory texts (e.g., Barnes, 1964; Karger & Bayah, 1965; Mundel, 1960). Researchers must define the purpose of the study and select behavioural categories to observe accordingly. They must then choose the level of accuracy required in the data and therefore the number of observations to be conducted, select a schedule of sampling and data collection, analyse the results, and prepare a report on the findings. Some texts have also recommended actively seeking the support of management in work sampling studies so that data collection is facilitated and the value of work sampling understood (Hansen, 1960; Niebel, 1967).
<table>
<thead>
<tr>
<th>Field</th>
<th>Category</th>
<th>Sub-approach</th>
<th>Other names</th>
<th>Purpose</th>
<th>Method</th>
<th>Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work study / operation</td>
<td>Methods design / study</td>
<td>Operation analysis</td>
<td></td>
<td>To identify the most efficient and fastest method of completing a task</td>
<td>Detailed study of each component of the task</td>
<td>Process charts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Motion study</td>
<td>Motion economy</td>
<td>To eliminate unnecessary operator movements</td>
<td>Analysis of body movements</td>
<td>Micromotion study</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Memomotion study</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time study</td>
<td></td>
<td>To determine how long a task should take to complete</td>
<td>Establishment of time standards</td>
<td>Stop watch / Marsto-chron</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Activity sampling</td>
<td>Ratio delay</td>
<td>To determine how people spend their time</td>
<td>Samples taken and proportions estimated</td>
<td>Direct observation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Work sampling</td>
<td></td>
<td></td>
<td>Time of Movements chart</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Random observation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Performance rating</td>
<td></td>
<td>To determine a reasonable time in which an operator is required to</td>
<td>Observe ‘average’ operator to establish ‘standard time’</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>complete a task</td>
<td></td>
<td>Rating system</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Job analysis</td>
<td>Job component</td>
<td>To identify the 'worth' of work of minimum standard</td>
<td>Wage determination</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Classification method</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Point system / ranking</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Factor comparison</td>
<td></td>
</tr>
</tbody>
</table>

*Table 1.*

Methods, Sub-approaches, Alternative Names, and Methods in the Field of Work Study.
SELECTION OF BEHAVIOURAL CATEGORIES

Prior to commencing work sampling data collection, the behavioural categories to be observed must be determined. The categories selected will be influenced by the purpose of the study (Heiland & Richardson, 1957). For example, if the goal of the study is to identify ways of increasing productivity, simply recording ‘productive’ and ‘unproductive’ categories may be sufficient. Precise and clear definitions for each category must be written (Barnes, 1964) and the inclusion of examples may be useful. Categories should be distinguishable and clear to assist observers to record accurately. In addition, categories should be exhaustive (encompassing all aspects of the job or process), and mutually exclusive (Kirwan & Ainsworth, 1992). Mutual exclusivity results in a requirement for decision criteria for situations in which an observed behaviour may fall into multiple categories. Mutually exclusive categories are desired for ease of recording and affected by the information required from the study. Due to the way the data are analysed (e.g., presence or absence of each behaviour at each observation), categories need not be mutually exclusive, and this may reduce the loss of information that may occur under mutually exclusive categories. However, it does not appear that methods of deriving categories have been explored in work sampling. Richardson and Pape (1982) suggested that the categories chosen should be easily recognisable in observations and written down so that observers may reference definitions and consistency between observers is facilitated.

In the hypothetical case study, the work sampling expert discussed with the management team the purpose of the study and why the study had been requested. The group also discussed the types of activities in which nurses were typically involved. In addition, the work sampling expert consulted with several senior emergency nurses and reviewed the literature on work sampling in hospital settings in order to develop a list of behaviours in which nurses were likely to be engaged during their shift. This list was further refined by the
expert reviewing nurse job descriptions and spending one shift shadowing a consenting nurse (and patients) to record the activities in which they were engaged.

DEFINITION OF BEHAVIOURAL CATEGORIES

Although traditional work sampling uses one list of mutually exclusive categories, multilevel task categories have been developed and used in more modern work sampling studies. Robinson (2010a) developed broad categories under which more specific activities were recorded for a more detailed data collection. He reported detailed techniques used to develop the categories including reviews of previous studies and analyses of the data collected by these studies, interviews with experts in the field, analyses of job descriptions and organisation policy documents, preliminary observations and data recording, a brainstorming exercise using cards to be sorted into categories and on-going feedback, and discussion with field experts. Although a thorough approach allowed modifications to be made to categories if needed, it was noted to be time-consuming (Robinson, 2010a). The use of pre-developed categories that have been piloted and tested for content validity in the area in which they apply may reduce this time-consuming step in designing a work sampling study. For example, Lo et al. (2007) adapted a list of categories developed by Overhage, Perkins, Tierney, and McDonald (2001) for studying physician behaviour.

In the example case, having selected behavioural categories, the work sampling expert then had developed a list of behaviours that were mutually exclusive and exhaustive (i.e., covered the full range of possible activities). These included paperwork and computer work, direct patient care, professional co-worker interaction, and breaks. Each category was defined clearly and approved by management prior to the study commencing.
DATA COLLECTION

Data are collected by noting down the category of behaviour in which the worker is engaged at each snap observation. Very few studies define ‘snap’ observations, however a ‘snap’ observation has been suggested to be between 5 and 10 s in duration (Jackson, 1972). Mobile positioning, in which the observer stands close enough to the worker to see what they are doing and follows them when they move, is often used (Burke et al., 2000). The observer must remain far away enough to avoid being intrusive. When more than one worker is being observed, observations can be conducted in a rotation. For example, Williams, Harris, and Turner-Stokes (2009) observed several nurses in turn, with each rotation of observations repeated every 5 min resulting in large amounts of data collected. Observing workers in rotation can be an efficient use of observer time because they are not required to wait for periods of time for the next observation time to be scheduled (Hansen, 1960).

Although it is desirable for observers to be able to see what the worker is doing, if there are instances where the observer can observe the general category but not the specifics (for example, the worker is writing but the observer cannot see whether they are filling out a form or writing a memo), Richardson and Pape (1982) suggested that it is acceptable for the observer simply to ask the worker. Although asking may facilitate more complete data collection, it may also influence the reactivity (potential change in worker behaviour in the presence of an observer). The objectivity of the data may also be compromised. Karger and Bayha (1965) stated that it is generally undesirable for the observer and worker to interact but that the worker should be aware of the purposes of the study and that they are being observed. Informing workers of the purposes of the study may increase reactivity regardless of interactions between observers and workers.

Barnes (1957) suggested that a new data form, tailored to the purposes and categories of the study, should be made for each study. The observer traditionally carries data sheets on
a clipboard and notes the observed activity at each observation point in time. Pelletier & Duffield (2003) suggested that a good data sheet should be easy to read and understand, contain clear boxes to mark, and denote sampling times. Some data collection forms also have spaces in which observers can write qualitative notes regarding what they have observed (e.g., Knickman, Lipkin, Finkler, Thompson, & Kiel, 1992). Qualitative notes can be useful when an observer is unsure about into which category a behaviour falls. If the categories are mutually exclusive, there must be decision criteria in place for observers to choose which category to record if the worker is performing several tasks at once (Pelletier & Duffield, 2003).

Some studies categorise behaviours generally as well as developing subcategories. Knickman et al. (1992) developed 67 coded subcategories which described what a physician was doing within broader categories. For example, under the general category of ‘writing’, the physician was coded as writing a note, a prescription, or filling out a form. No data on how a more complex data recording method affects accuracy or ease of recording were reported in this study. Some fields have developed occupation-specific tools for use in work sampling. For example, Hoffman, Tasota, Scharfenberg, Zullo, and Donahoe (2003) used the Clinician Activities Tool to observe and categorise the behaviour of acute care nurses and physicians. The tool is comprised of 42 activities under nine subcategories and three major categories. Similarly, the Posture, Activity, Tools and Handling tool (PATH) has been developed to perform work sampling in highway construction work settings (Buchholz, Paquet, Punnett, Lee, & Moir, 1996). The PATH tool is customised for the work site being studied and fixed-interval recordings of the PATH category as well as the job being done is recorded. Buchholz et al. found the PATH tool to be both reliable and valid in their pilot study. Traditional methods of work sampling use pen-and-paper recording, however there have been many variations reported in data collection for work sampling studies.
The development of electronic data recording in work sampling studies has produced some variation in the way that data are collected. Quach et al. (2011) reported the use of tablet computers programmed with electronic data sheets for the collection of work sampling data. The tablet computer also contained a timer that ran alongside the data sheet, permitting continuous observation of immunization staff collecting patient information. Similarly, observers in a study by Lo et al. (2007) recorded mutually exclusive categories of physician behaviour on touch-screen computers. Electronic data recording methods such as these facilitate continuous recording and may allow more data to be collected than the traditional pen-and-paper observation forms. However, there do not appear to have been any studies in the work sampling literature that have compared the data obtained by electronic means to traditional work sampling studies. Therefore, it is not known if electronic recording is easier, more efficient, more effective, or susceptible to error. Caughy and Chang (1998) used barcode scanners for nurses to self-record activities throughout the day. Nurses scanned bar codes that recorded the activity being performed, as well as the time. Therefore, continuous data were recorded without the need for observers in an efficient way. Although more quantitative than self-report, these data still have the potential for bias or untruthful recording. In addition, there may be some work settings where confidentiality issues preclude the presence of an observer (Stevenson, Caverly, Srebnick, & Hendryx, 1999).

There have been some comparison studies comparing the use of direct observations or other objective measures with self-reports. Self-reports may be recorded in a number of ways. For example, Stevenson et al. (1999) gave staff in a mental health setting a pager that vibrated at random times. When the pager vibrated, the staff noted what they were doing and the time in a log. However, when Ampt, Westbrook, Creswick, and Mallocki (2007) compared nurses self-reports to direct observations, they found that self-reports were not reliable. Ampt et al. did not, however, keep the number of data points constant between
sampling methods, nor did they compare either method to true values. Therefore, it can only be claimed that the two methods did not produce consistent results, not that one was more reliable than the other. Oddone, Guarisco, and Simel (1993) asked staff in a general medical service to complete a coded book when their beepers vibrated. These data were compared to estimates that the staff made (prior to the coded book recordings) of the proportion of time that they spent engaging in each activity. Oddone et al. found that staff made poor estimates, mostly greatly overestimating the time they spent in each category. Burke et al. (2000) compared direct reports to continuous recording by an independent observer, finding a low level of correspondence between the two. Oddone et al., found that self-reports greatly overestimated the time spent in each category. As the two data sets were collected one after the other, they could therefore only be expected to correspond if the behavioural categories of the nurses studied varied within each data collection period in a consistent way. Burke et al. suggested that fewer and clearly defined categories were required in self-reporting to minimise inconsistencies between responding across participants. Generally, self-report methods are considered to be biased and when self-reports are recorded retrospectively, they can be inaccurate due to limitations in recollection (Robinson, 2010a).

It is possible to record activities and code categories from videos (Segall & Kotzan, 1979), but this can be time-consuming, particularly if the video camera is unable to be left in one position and must be operated by a person (therefore doubling the amount of time involved in data recording). Niebel (1967) suggested the use of a random activity analysis camera, whereby recorded data can be analysed frame by frame. He argued that this method reduces reactivity, and reported a study in which the proportion of time spent not working was recorded as larger when analysed from video than when directly observed in real time. The results of the study suggested that workers were changing their behaviour in the presence of an observer. Barnes (1957) also recommended the use of a camera and a timer on the
screen to permit the human recorder to make snap observations to simulate direct observation work sampling. The advantage of data taken from a permanent record such as a video is that the video can viewed multiple times by multiple people, therefore facilitating reliability checks.

Training of observers is conducted in some studies and the type of person used to observe varies. For example, Lo et al. (2007) trained research assistants by having them directly observe people in the setting in which they were to collect data, and by providing explicit training on the data recording tool (on touch-screen computers). Bartholomeyczik and Hunstein (2004) trained observers using videoed examples of the behaviour categories and practice observation sessions in settings similar to those in which the study was to be conducted. Caughy and Chang (1998) trained nurses in a hospital to be data coordinators. The nurses were taught to troubleshoot problems with the electronic equipment and the operation and maintenance of the equipment. The coordinators were also provided with didactic information and supervision, however Caughy and Chang suggested that coordinators would have also benefitted from more extensive training, a training video, and a list of frequently asked questions to which to refer.

Although practitioners may be preferred observers because they are familiar with the work to be observed, administrative staff can be used. For example, Quist (1992) used administrative staff in an effort to increase objectivity in the measurement. Lindenmeyer & Chisman (1980) developed a computer game to teach industrial engineering students the concepts, methods and mathematical components of work sampling. Although developed to motivate students and enhance understanding of work sampling, Lindenmeyer and Chisman suggested that their computer program could also be used in work settings as a simple and effective way to teach work sampling methods. The development of work sampling
programs may lead to an efficient and effective way to train multiple observers, and therefore reduce the time required to prepare for a work sampling study.

When multiple observers are used, inter-observer agreement is often reported. Pelletier and Duffield (2003) suggested that reliability checks between observers should be conducted randomly throughout a study, and should be above 90% to be considered satisfactory. However, Pelletier and Duffield also suggested that if only one observer is recording data for a study, that inter-observer reliability is not necessary. Although it is the case that inter-observer checks are not possible with one observer, no alternative method of conducting checks of the reliability or accuracy of the data collected were suggested by Pelletier and Duffield. It could be argued to be unsatisfactory to dismiss reliability checks as not possible and therefore unnecessary. The ability of observers to identify and correctly code categories will affect the reliability of the data (Zheng, Guo, & Hanauer, 2011) and thus this should be evaluated. Williams et al. (2009) used two observers to record data simultaneously but not independently (as they argued that independent recording was not possible in this setting). The two observers were used to refine and discuss the measurement system.

*In the nurse work sampling example, observations were to be conducted by a team of observers across all shifts (24 hours per day), observations were scheduled for every 5 min (24 hr = 1,440 min divided by 292 observations per day). A group of 10 nurses were selected to participate in the study and observations were conducted cyclically (e.g., the first nurse was observed on the first observation, the second on the second etc.), resulting in each nurse being observed every 50 min. Observers had five min in which to locate the next nurse to observe.*
WORK SAMPLING EQUATION

Although Karger and Bayha (1965) suggested that many researchers use the Rule of 1,000 in work sampling research (1,000 observations will be sufficient to capture representative samples in most studies), work sampling has a mathematical basis that provides a tool for calculating the required number of observations. An observation consists of looking momentarily at a worker at a predetermined time and recording the behaviour in which they are engaging. The underlying statistical concepts of work sampling are those of random sampling, and the normal and binomial distribution curves. It is assumed that a sample of what people are doing will have the same distribution as what they are doing across a whole day, week, or other time-of-interest (laws of probability). Although the normal distribution curve has been presented to explain the statistical basis of work sampling (Barnes, 1957), the data collected in work sampling studies are attribute data. Attribute data are those in which each datum is reported as belonging to one of a number of categories, resulting in a binomial distribution (e.g., ‘yes-no’ data). A binomial distribution approaches a normal distribution when the number of independent trials or observations approaches infinity (Wetherill & Brown, 1991). As work sampling often utilises large numbers of samples (the term ‘sample’ refers to a single observation in the context of work sampling), the normal distribution is sufficiently close to the binomial distribution (Niebel, 1967). Although work sampling typically measures more than two categories of work behaviour, each category is analysed by determining at each observation point whether the subject is engaged in this category (i.e., occurrence or non-occurrence is recorded). The binomial distribution underpinning work sampling data thus determines the error in the sampling, as well as the total number of observations required to obtain a sample that will be accurate at the selected level of accuracy (Barnes, 1957).
The equation used in work sampling that includes both accuracy and the number of observations required has a basis in the equation for standard deviation (the average distance of each data point from the mean); a measure of variability. The equation is:

\[ \sigma_p = \sqrt{\frac{p(1-p)}{N}} \]  

(1)

\( \sigma_p \) represents one standard deviation (denoted by sigma), \( p \) represents the proportion or percentage occurrence of a behavioural category or event (expressed as a decimal), and \( N \) represents the number of observations in a sample (Mundel, 1960). The equation for one standard deviation shows that 68.27% of samples of size \( N \) will produce a \( p \) value within one standard deviation of the mean (also referred to as a confidence interval). More simply, confidence intervals of 68.27% indicate that 68.27 samples out of 100 will reflect what is truly occurring (Denton, 1987). Should a higher level of certainty that the sample will fall within true values be required, a confidence level of 95% can be used (two standard deviations). The equation becomes:

\[ \sigma_p = 2 \sqrt{\frac{p(1-p)}{N}} \]  

(2)

Barnes (1964) suggested that 95% confidence intervals are typically satisfactory for work sampling, and although some of the work sampling texts vary in the equations reported, many appear to use Equation 2. The use of Equation 2 permits two approaches in determining sample size and accuracy. It is possible to determine the accuracy of the results obtained based on the number of samples conducted in the study. It is also possible to calculate the number of observations that must be conducted in order to achieve a sample within the desired level of accuracy (Richardson & Pape, 1982). In order to calculate the number of observations, the equation is transformed in the following manner:
\[ Sp = 2 \sqrt{\frac{p(1-p)}{N}} \]

\[ S^2p^2 = 4 \frac{p(1-p)}{N} \]

\[ N = \frac{4p(1-p)}{S^2p^2} \]

\[ N = \frac{4(1-p)}{S^2p} \] \( (3) \)

Sp in Equation 3 represents the desired accuracy (replaces the notation for standard deviation in the previous equations). For example, to obtain results with a 95% confidence interval and within ± 5% accuracy, Sp is .05. The following is an example of the use of the equation to obtain, within ± 5% accuracy, the required number of observations for a behaviour that occurs for an estimated .1 (10%) of the time.

\[ N = \frac{4(1-p)}{S^2p} \]

\[ N = \frac{4(.1)}{.05^2 \times .1} \]

\[ N = \frac{4(.9)}{.0025 \times .1} \]

\[ N = 14,400 \]

Although the practicality of conducting this number of observations would be determined by the duration of the study, 14,400 observations is a larger number to schedule and conduct.

The number of observations required increases greatly when p is smaller than .1. For example, Equation 3 calculates 30,400 observations are required when p is .05 (5%) and 158,400 when p is .01 (1%). These numbers are large and may be impractical to conduct in some settings. The practicality of large numbers of observations is likely to be influenced by the duration of the study, which is determined by factors such as cost and resources.
Figure 3. The number of observations required by Equation 3 for ±1% to ±5% relative accuracy, as a function of the estimate of the proportion of time spent in the activity of interest (p).

Equation 3 calculates the number of observations required with relative accuracy (Figure 3). Relative accuracy of 5% would be ±5% of the value of p. For example, relative accuracy of 5% for a behaviour that occurs .1 of the time would indicate that the observation values fall within .095 and .105; the relative accuracy on either side of the .1 value is 5% of that .1 (±5%). However, when the values of p are much larger, e.g., .9, the use of relative error produces a much larger range in which the accuracy of obtained values fall; 5% of .9 is .045, therefore results are accurate with .855 and .945. Barnes (1957) suggested that absolute error may be considered more acceptable and more logical (as it would be unlikely that errors of different values would be acceptable for different p values). Therefore, another version of Equation 3 can be used that uses absolute error (the error values are constant on either side of the p value, regardless of the p value). For example, under this equation, results would be accurate within .05 and .15 for a p value of .1 and .85 and .95 for a p value of .90.
\[ N = \frac{4p(1-p)}{S^2} \] (4)

In Equation 4, all of the variables remain the same as in Equation 3. Equation 4 was inferred from the data presented in a table by Barnes (1957) as an alternative to Equation 3. Barnes’ table showed that when Equation 4 is used, the number of required observations is greatest when \( p \) is estimated to be .5, with an equal number of observations required for .1 and .9, .2 and .8 etc. (Figure 4). It should be noted that 95% confidence intervals are also used in Equation 4.

**Figure 4.** The number of observations required by Equation 4 for \( \pm 1\% \) to \( \pm 5\% \) absolute accuracy (S), as a function of the estimate of the proportion of time spent in the activity of interest (\( p \)).
The number of required observations calculated with the Equation 4 is much less than required by Equation 3 (Table 2). The largest number of required observations under the absolute error equation is calculated when absolute accuracy is ± 1%; 10,000 observations are required when \( p \) is .5. In contrast, the highest number of required observations is calculated using the relative error equation when desired accuracy is ± 1% and \( p \) is .05; 760,000 observations are required (over 200 hours). The distribution of observations over time is determined by the method of sampling selected and the available time in which to conduct the study. The greatest differences in the equations occur when \( p \) is small (e.g., the activity or behaviour occurs for a small percentage of the time). However, when \( p \) is large, the equations produce much more similar numbers of required observations. For example, with ± 1% accuracy, the relative error equation produces 2,105 observations and the absolute equation produces 1,900 observations (Table 2). As in all sampling procedures, a greater number of samples is considered to more closely approximate whole time-of-interest values than smaller sample sizes (Brisley, 1992).

In the case example, the work sampling equation was used to calculate how many observations were to be conducted over the one month study in order to obtain a representative sample of how nurses were spending their time. The expert was able to choose between absolute or relative acceptable error. Absolute error of ± 5% indicates that the results are accurate within .05 of the obtained proportion value. For example, when nurses are found to spend .12 of their time completing paperwork, the true value is predicted to be between .07 and .19. Relative error indicates that the results are accurate within ± 5% of the obtained proportion value. For example, relative error for an obtained value of .12 would be accurate within .11 and .13. Relative error is therefore more stringent. The work sampling expert selected relative error as this resulted in an acceptable degree of error for the purposes of the study. Additionally, the larger number of observations required for relative
error (compared to absolute) was manageable over the month of observations. The expert calculated that 9,067 observations were required to obtain an acceptable sample and divided the days of the study (31) by the number of observations required to calculate how many observations to conduct each day (292).
Table 2.  
Number of Observations Calculated under Equations 3 and 4 (Relative Error (RE) and Absolute Error (AE), respectively) for a Range of Desired Accuracies (± 1% to ± 5%) and PO Values.

<table>
<thead>
<tr>
<th>Percentage Occurrence of Activity (PO)</th>
<th>± 1%</th>
<th>± 2%</th>
<th>± 3%</th>
<th>± 4%</th>
<th>± 5%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RE</td>
<td>AE</td>
<td>RE</td>
<td>AE</td>
<td>RE</td>
</tr>
<tr>
<td>5%</td>
<td>760,000</td>
<td>1,900</td>
<td>190,000</td>
<td>475</td>
<td>84,444</td>
</tr>
<tr>
<td>10%</td>
<td>360,000</td>
<td>3,600</td>
<td>90,000</td>
<td>900</td>
<td>40,000</td>
</tr>
<tr>
<td>15%</td>
<td>226,667</td>
<td>5,100</td>
<td>56,667</td>
<td>1,275</td>
<td>25,185</td>
</tr>
<tr>
<td>20%</td>
<td>160,000</td>
<td>6,400</td>
<td>40,000</td>
<td>1,600</td>
<td>17,778</td>
</tr>
<tr>
<td>25%</td>
<td>120,000</td>
<td>7,500</td>
<td>30,000</td>
<td>1,875</td>
<td>13,333</td>
</tr>
<tr>
<td>30%</td>
<td>93,333</td>
<td>8,400</td>
<td>23,333</td>
<td>2,100</td>
<td>10,370</td>
</tr>
<tr>
<td>35%</td>
<td>74,286</td>
<td>9,100</td>
<td>18,571</td>
<td>2,275</td>
<td>8,254</td>
</tr>
<tr>
<td>40%</td>
<td>60,000</td>
<td>9,600</td>
<td>15,000</td>
<td>2,400</td>
<td>6,667</td>
</tr>
<tr>
<td>45%</td>
<td>48,889</td>
<td>9,900</td>
<td>12,222</td>
<td>2,475</td>
<td>5,432</td>
</tr>
<tr>
<td>50%</td>
<td>40,000</td>
<td>10,000</td>
<td>10,000</td>
<td>2,500</td>
<td>4,444</td>
</tr>
<tr>
<td>55%</td>
<td>32,727</td>
<td>9,900</td>
<td>8,182</td>
<td>2,475</td>
<td>3,636</td>
</tr>
<tr>
<td>60%</td>
<td>26,667</td>
<td>9,600</td>
<td>6,667</td>
<td>2,400</td>
<td>2,963</td>
</tr>
<tr>
<td>65%</td>
<td>21,538</td>
<td>9,100</td>
<td>5,385</td>
<td>2,275</td>
<td>2,393</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>70%</td>
<td>17,143</td>
<td>8,400</td>
<td>4,286</td>
<td>2,100</td>
<td>1,905</td>
</tr>
<tr>
<td>75%</td>
<td>13,333</td>
<td>7,500</td>
<td>3,333</td>
<td>1,875</td>
<td>1,481</td>
</tr>
<tr>
<td>80%</td>
<td>10,000</td>
<td>6,400</td>
<td>2,500</td>
<td>1,600</td>
<td>1,111</td>
</tr>
<tr>
<td>85%</td>
<td>7,059</td>
<td>5,100</td>
<td>1,765</td>
<td>1,275</td>
<td>784</td>
</tr>
<tr>
<td>90%</td>
<td>4,444</td>
<td>3,600</td>
<td>1,111</td>
<td>900</td>
<td>494</td>
</tr>
<tr>
<td>95%</td>
<td>2,105</td>
<td>1,900</td>
<td>526</td>
<td>475</td>
<td>234</td>
</tr>
</tbody>
</table>
NOMOGRAPHS

An alternative and faster method of calculating the number of observations required is the use of a nomograph or alignment chart (Barnes, 1957). A nomograph is a diagram of graduated lines, each representing a variable, from which values of an unknown variable can be read when the values of the other variables are known (Allcock & Jones, 1950) (for example, Figure 5). Figure 5 represents the absolute accuracy work sampling equation as a diagram comprised of a scale for the value of $p$ (proportion of time spent in a behavioural category), a precision scale (referred to as accuracy in the equation), and a scale representing the number of required observations.

Figure 5. Nomograph showing the number of required observations calculated (Equation 4) from a $p$ value of 0.8 or 0.2 and required accuracy of ± 4% for a confidence level of 95%. From “Work sampling without formulas” by J. M. Allderige, 1954, Factory Management & Maintenance, 112, p.136. Copyright McGraw-Hill Publishing.
In order to use Figure 5 to determine the number of observations required for a behaviour assumed to occur .8 of the time with absolute accuracy of ± 4%, a straight line can be drawn that connects these two points and intersects the third scale (number of observations). The number of observations can then read from the third scale at the intersect (400 observations). Logarithmic scales are most commonly used in the construction of nomographs, with the distances between points on the scale calculated by the following equation (Allcock & Jones, 1950):

\[ \mu (\log y - \log x) = \text{length of scale (inches or centimetres)} \]  

where \( \mu \) represents the scale factor, \( y \) the highest value on the scale, \( x \) the lowest value on the scale, and the length of the scale on the page may be selected fairly arbitrarily. Equation 5 calculates the scale factor, which is then multiplied by the difference between logs of the values on the scale to find the distance (on the page) between them. For example, for a scale that is 15 centimetres in length, and has 10,000 and 100 as end points, \( \mu \) is calculated to be 5:

\[ \mu (\log 10,000 - \log 100) = 15 \text{cm} \]

\[ \mu (4 - 1) = 15 \text{cm} \]

\[ \mu = 5 \text{cm} \]

Therefore, the distance between 4,000 and 3,000, and 3,000 and 2,000 on the scale is calculated to be:

\[ \text{distance} = 5 \times (\log 4,000 - \log 3,000) \text{ and distance} = 5 \times (\log 3,000 - \log 2,000) \]

\[ \text{distance} = 5 \times (3.60 - 3.47) \text{ and distance} = 5 \times (3.47 - 3.30) \]

\[ \text{distance} = 0.62 \text{cm or 6.2mm and distance} = 0.88 \text{cm or 8.8mm} \]

Equation 5 permits a researcher to calculate the values of and construct a nomograph. Two nomographs for use in work sampling were provided by Allderidge (1954) (Figure 5) and Moskowitz (1965). On these nomographs, all three scales (the value of \( p \), the accuracy
required, and the number of observations) are logarithmic and permit a researcher to use two
known values to determine the third.

To use any nomograph in work sampling, the desired level of accuracy and estimated
value of $p$ are located on the scales and joined with a straight line. The line is then extended
to intersect the scale that determines the number of observations. By the same method, the
nomograph can be used to determine the level of accuracy given the number of the
observations conducted, and the confidence intervals given the level of accuracy and the
number of observations (Richardson & Pape, 1982). Richardson and Pape reported a
nomograph with four scales (confidence intervals, percentage of occurrence, number of
observations, and accuracy), whereas others have reported nomographs that have only three
scales and assume confidence intervals of two sigma limits (95%) (Barnes, 1957; Brisley,

When the number of required observations within the desired accuracy is calculated,
the number of observations is divided amongst the number of days over which the study is to
be conducted. Barnes (1957) suggested that the number of days over which the study is
carried out must be large enough to capture all variability in activities. For example, if a
month is considered to encompass all of the variability in activities affected by factors, such
as end of the month stock-take and monthly meetings, then this will be a sufficient period
over which to distribute observations. Barnes also suggested that the longer the period of
time over which observations are spread, the easier it will be to obtain a representative
sample (relative to a shorter study period with the same number of observations). The
duration of the study will also be affected by the sampling schedule used. The number of
observations to be conducted per day is determined prior to the commencement of the study.
After the determined number of observations is conducted on the first day, an observed value
of the proportion of time spent engaged in an activity ($p$) is obtained. The obtained $p$ value is
then entered back into Equation 3 or 4, and a new number of required observations is calculated. Morris (1969) suggested that if the new calculated number of observations required is greater than the original calculation, the additional observations required are added and distributed across the remaining days of the study. However, if the required number of observations is lower than the number originally calculated, Morris recommended retaining the original number because this will provide a larger sample size (assumed to be a closer approximation to overall durations of the categories). An alternative method of using the equations would be to determine the number of observations able to be conducted, and then solving the equation to find the error of the sample size (Hansen, 1960). Calculating obtained error may be useful when the number of observations able to be conducted is restricted by the setting (for example, the availability of observers or resources).

Typically, the observed proportions of time spent in the categories of work are assumed to be indicative of overall proportions of time spent in these categories, not just the proportion of time spent across the days being studied. Therefore, the true value of these proportions sought through sampling is an average of variations in this proportion across days (Richardson & Pape, 1982). In calculating random-day effects (daily variance estimates), variance can be analysed day-by-day, and daily fluctuations can be ignored if shown to be small (Richardson & Pape, 1982). Analysing variance daily may help the researcher to adjust the number of observations to be conducted (e.g., the number may increase if the variance between daily data is found to be high).
Burr (2005) suggested that Statistical Process Control (SPC), from which control charts are taken, is a relatively simple statistical method of assessing the variability of data to assist in decision making. SPC, developed by Walter A. Shewart in the 1930s and famously used by W. Edwards Deming to revolutionise production lines in Japanese companies in the 1980s and 1990s, was originally used to increase the yield or quality in manufacturing processes (Wetherill & Brown, 1991). SPC plots data describing a process (e.g., attribute data for a product from a factory is either acceptable or defective), and uses statistical analyses to identify any data points that deviate from others (Hopkins, 1995). Underpinned by the assumption that all processes (in the context of industrial production lines) have variable outputs, SPC aims to determine whether variability is normal (e.g., due to inconsistent raw materials, variation in workers) or abnormal (e.g., broken equipment). Abnormal variability requires a change to the system, as it shows the process to be out of statistical control (Mainstone & Levi, 1988). In the context of work sampling, these highly variable and abnormal data points are discarded as erroneous or unimportant or analysed.

Control charts can be used to plot the daily results of observations to determine whether these obtained values fall within the control limits of the study (Barnes, 1964). Control limits are calculated using the equation for standard deviation and plotted as broken horizontal lines on a graph. Control limits thus show the levels (both above and below the average) at which the data deviate from the mean. Any values that are highly variable and fall outside control limits are considered to be outliers and caused by an identifiable factor. These data points are discarded from the study and observations are conducted for an additional day (Barnes, 1957).

A control chart plots the means of observation samples in temporal order, and it is assumed that these samples were collected in temporal proximity (Wetherill & Brown, 1991).
Control charts can be produced for both continuous variables such as height or volume, and for attribute data such as the data collected in work sampling (Mainstone & Levi, 1988). In order to construct a control chart for a work sampling study, a mean $p$ value (proportion of occurrence for the category of interest) is calculated and plotted as a central line intersecting the $y$ axis. Burr (2005) suggested that for attribute data, the value of $p$ may be taken from previously collected data or a desired value may be used in the context of production processes. Heiland and Richardson (1957), however, suggested that the value of $p$ could be calculated by taking the mean of the first four or five observations. For shorter studies, this may not be possible if the purpose of the control chart is to track obtained data and adjust control limits accordingly. Control limits are calculated, usually assuming a three-sigma limit (three chances in 1,000 that data lying outside these limits will be due to chance) (Barnes, 1964). It is also possible and perhaps useful to calculate the limits at one or two sigma, depending on the confidence intervals required (Brisley, 1992). Control limits for a central mean $p$ value of .1 at a three-sigma control limit are as follows:

$$\text{limits} = p \pm 3 \sqrt{\frac{p(1-p)}{n}}$$  \hspace{1cm} (6)

$$\text{limits} = .1 \pm 3 \sqrt{\frac{.1(1-.1)}{2880}}$$

$$\text{limits} = .1 \pm .017$$

$$= .117 \text{ and } .083$$

$$= 11.7\% \text{ and } 8.3\%$$

When calculating control limits, $n$ is the number of daily observations (calculated by dividing the number of required observations obtained from the work sampling equation by the number of days over which the study is to be conducted). In the example calculation using Equation 6, $n$ was calculated by dividing the number of observations determined in the previous worked example of Equation 3 by five days ($14,400 / 5 = 2,880$). The $p$ value used
was the predicted $p$ value also used in the previous worked example.\footnote{A nomograph can be used as an alternative to the equation to calculate control limits. Such nomographs are comprised of a sub-sample scale ($n$), a proportion of occurrence scale ($p$), and a plus or minus control limits scale (presented as a percentage) (Barnes, 1964).} Figures 4 and 5 are two theoretical control charts using the values calculated above and showing all the required elements of such a chart. The data shown are for the purposes of illustration and were generated by randomly selecting nine data points between the control limits calculated for $p = .1$, and two data points between the upper control limit and .14. Figure 6 applies constant control limits at a three-sigma level (based on the initial $p$ value). Figure 7 shows adjusted control limits for each data point. Adjusted control limits were calculated by using each simulated ‘obtained’ $p$ value (from observation days 1 to 11) to calculate corresponding control limits.

![Figure 6](image)

*Figure 6.* An example control chart with constant control limits (three-sigma) constructed from theoretical data.
Figure 7. An example control chart with changing control limits (three-sigma) constructed from theoretical data.

Control charts assume that samples taken on subsequent days were comprised of an equal number of observations. It is therefore recommended that when this is not the case, control limits be calculated for each sample. An average $n$ may be used to calculate control limits when all sample sizes lie within $\pm 5\%$ of the average $n$ (Burr, 2005). In interpreting a control chart, data points that lie outside control limits are assumed to be due to an assignable cause may therefore be discarded. However, Wheeler and Chambers (1992) suggested three other patterns of data that may indicate an assignable cause. First, a pattern of at least eight consecutive points either above or below the central line. Second, at least 4 out of 5 consecutive points fall on the same side of the central line and outside the one-sigma limit. Third, at least two out of three consecutive points fall on the same side of the central line and outside the two-sigma limit. In addition, Burr (2005) suggested that assignable causes may be identified when data points alternate between high and low or when there is a noted trend of at least eight consecutive points. Therefore, control charts provide a decision-making tool
that incorporates the use of rules (Pfadt & Wheeler, 1995). Brisley (1992) also suggested that control charts can be used to determine empirically how long the study should be by calculating the control limits at a chosen confidence level for the cumulative number of observations. The study is terminated when the required control is demonstrated. Control charts can also be constructed across observations (each observation is treated as a sample). Constructing control charts across observations provides within-session data (e.g., within a day), and may be used to verify that the samples have been distributed randomly (if that was the sampling method used) (Heiland & Richardson, 1957). Control charts can also be used to detect patterns across the day or people when sampling is cyclic across workers.

There are some limitations to the conclusions drawn from control charts. Although a process can be determined to be within statistical control (i.e., stable and predictable), no inferences can be made regarding the acceptability of the products of the process (Mainstone & Levi, 1988). For example, if data are collected on the number of car parts produced by a production line on each day of observation, a control chart could be used to determine whether the number of parts produced is consistent, but could not be used to determine how many of those parts were faulty. Similarly, a control chart may indicate that nurses spend a similar proportion of each day of the study completing paperwork, however cannot be used either to evaluate the quality of the paperwork or to determine whether the amount of time spent conducting paperwork is too much or too little. In addition, because the plotted means are a result of sampling, which produces variability in obtained data, Type I errors (in which a false positive is identified), and Type II errors (in which an important change has been missed) are possible (Wetherill & Brown, 1991).

*Variations in the proportion of time spent in each category on certain days were tracked using a control chart and discussed with the management. For example, the proportion of time spent on breaks was much lower than other days on one particular*
Saturday. The variation was evidenced by the plotted proportion value for breaks lying outside the lower three-sigma limit on the control chart. Upon comparison with other data collected by hospital records, it was determined that the number of in-patients to the emergency department that day was much higher than usual due to poor weather and increased road traffic accidents.

**INITIAL ESTIMATES OF $p$**

Through the application of work sampling equations, it becomes apparent that both the work sampling equations to calculate the number of observations required to complete the study, and the equation used to calculate control limits require a known value of $p$ (proportion of time during which a particular activity occurs). Although the value of $p$ used in the control limits equation may be calculated by taking the mean of the first few days of observation, an initial $p$ value is required to commence the study with a determined number of required observations. Estimates of $p$ to be inserted in the equation can either be estimated (usually from the results of previous studies or reports), or determined through preliminary observations (Barnes, 1957). Hansen (1960) suggested that between 100 and 200 observations should be conducted in order to provide an estimate of $p$. Additionally, it is likely that a study will include more than two categories of activity to be observed, and therefore the observer must choose which $p$ value will be used in calculations. Richardson and Pape (1982) stated that one category may be of particular interest and so the corresponding $p$ value is selected. Alternatively, observers may choose the smallest estimated $p$ value for the largest number of observations and thus the most conservative sample (this only applies under the relative equation as the greatest number of observations will be calculated with a $p$ value of near .5 under the absolute equation). For example, Handa and Abdalla (1989) used Equation 4 and the most stringent number of observations (i.e., they
chose \( p \) to be .5 or 50%), without making any estimates of true values of \( p \). As the number of required observations is recalculated based on the data collected each day of the study in work sampling, the number of required observations is correct with new \( p \) values. However, a much greater number of observations than originally planned will be required if \( p \) greatly underestimates true values, and an observer may find that they have completed all required observations in the first day if the initial \( p \) value greatly overestimated true values. Under such circumstances, the observer may choose to conduct at least one more day of observation as the distribution of observations may affect the representativeness of the data collected (i.e., data may not reflect true values when collected in one long session).

In the case example, the work sampling expert conducted a preliminary study for one day prior to the main study. During this time, the expert conducted 100 observations spread across one day shift and one night shift in the emergency department (shifts as defined by nurses’ rosters). To conduct the observations, the expert used a list of 100 pre-determined observation times (in min) to locate the nurse being observed and record what they were doing at that time. At the end of the preliminary observations, the proportion of observations in which each behaviour was observed was calculated.

SELECTING OBSERVATION TIMES

Introductory texts on work sampling (e.g., Barnes, 1957; Hansen, 1960; Niebel, 1967) provide random number tables for use in selecting observation times. When the number of observations to be conducted each day is calculated, the subsequent number of nine-digit random numbers is selected from the table and written down as a schedule of times. For example, selecting the number 171,067 would schedule two observations, one at 1.43 p.m. and one at 12.40 a.m. (the second, third, fifth, and sixth digits are transferred from decimals to minutes: .71 of 60 min is 43 min) (Barnes, 1957). Any times outside the time of interest
(for example 12.40 a.m.), are discarded. The use of random number tables to determine random observation times both reduces potential reactivity in those being observed (who may come to predict when the observer will arrive), and may reduce the error in the sampling (Karger & Bayha, 1965). Heiland and Richardson (1957) discussed two types of error that may arise in work sampling which may account for differences between the results of the sample and true values; systematic error and random error. Systematic error (for example, a machine that is in operation between quarter past and quarter to of every hour is recorded as operating 1.0 of the time because observations are always made at half past the hour) can be reduced through randomising observations. Random errors are due to chance and are more difficult to reduce. Further considerations in choosing sampling times may include the number of observers, the number of people to be observed, the locations of the observations, and the distance between them (Heiland & Richardson, 1957). All factors should be considered and a schedule of observations developed prior to the commencement of the study.

Traditional work sampling uses ‘snap’ (momentary) observations of activities to predict overall durations of activities, which can be more cost effective than continuous observations (Finkler, Knickman, Hendrickson, Lipkin & Thompson, 1993). However, continuous observations can be considered samples when conducted for only a portion of the full time-of-interest. The use of automated or computerised data recording methods facilitates the collection of continuous observations (Ho & Pape, 2001), and has been used in modern studies.

Pape (1988) identified three methods of selecting observation times, all of which use stratification (the distribution of observations across days). Sampling is either random (observations are conducted at unequal intervals), or fixed (observations are conducted at intervals and equal duration of time apart). Restricted random sampling uses a random
number table to select observation times (the distribution of which is random). In systematic random sampling, the duration of the day (in minutes) is divided by the number of observations required to be sampled for the day. Observations are conducted across the whole day and the time between observations is kept constant (i.e., fixed intervals). Stratified non-continuous random sampling combines random and fixed sampling. The day in minutes is divided by the number of observations required to determine an interval duration (fixed), and a random number table is used to select a random observation time within each interval (random). Stratified continuous random sampling is a variation of stratified non-continuous random sampling, can be used to observe several workers, and requires the observer to make continuous observation rounds. To randomise the observations, the starting point or direction of the rounds is changed each day. Figures such as Figure 8 are presented in texts such as Currie (1963) to illustrate the different ways to schedule observation times.
Figure 8. The scheduling of observations under three momentary time sampling methods. Black blocks represent the duration of behaviour A and white blocks the duration of behaviour B. Observation times are denoted by vertical black arrows and the black vertical lines represent the division of the session into equal intervals for non-continuous random sampling.

There appear to be few studies that have tested directly the representativeness of data collected by work sampling methods, but there have been some studies comparing the result obtained under different sampling methods. Dickson’s (1978) study of pharmacy work compared the data collected by fixed-time sampling and random sampling, and found no difference between the two. However, it does not appear that the data collected by either method were compared to true values or continuous records. Therefore, although the two
sampling methods may have produced similar results, neither may have produced results representative of the actual or overall durations of activities. Although studies and texts claim that random sampling produces less bias than fixed-time sampling (e.g., Oddone & Simel, 1994), Finkler et al. (1993) suggested that for work that is not cyclic (e.g., particular activities are not performed at the same time every hour, two hours etc.), bias from fixed-time sampling maybe negligible. Similarly, Sittig (1993) suggested using fixed-interval sampling when the distribution of activities is random and random sampling when it is not. Fixed-time sampling may be easier to implement in practical settings because less time and effort are required to schedule observations (e.g., in comparison to using random number tables).

Liao and Pape (1996) suggested that fixed and random sampling are equally useful when behavioural categories occur randomly but argued that fixed-interval sampling is useful when the process under observation moves between categories in a random pattern. Liao and Pape explained the ‘random pattern’ in the context of the study period. They argued that the purpose of work sampling is to make inferences about the durations of behaviours outside the period being studied (otherwise it may be best to record continuous data for the whole time-of-interest), and therefore the random pattern observed in the data is assumed to repeat across time. In an effort to produce a model that enables \( p \) (the proportion of a behavioural category) to predict variability in the behavioural category outside of the study period, Pape (1993) used an alternating Poisson process (APP) to demonstrate the most efficient spacing between observations to obtain useful samples. The effect of this spacing on the randomisation of samples was also able to be quantified. Pape, through mathematical analysis, considered the ‘cost’ of reducing a study duration through the effect on how many more observations would then be required and thus the reduction in time between observations. Sampling efficiency under the APP model was shown to increase as the delay (and therefore the randomness) between observations was increased, however fixed-interval
sampling was much more efficient than random sampling (both overall and even more so at smaller delays between observations) (Pape, 1993). Pape therefore concluded that randomising observations is costly both in terms of the number of observations and duration of study required to achieve required results, but that using delayed sampling may be useful. Delayed sampling arranges a fixed time after each observation during which no further observations may be scheduled. After the delay, the next observation is scheduled after a second delay of random duration.

Liao and Pape (1996) recognised the disadvantages of the delayed-sampling method that it arranges no observations directly after an observation (a potential loss of information), and that it may be difficult to implement. As an alternative, they suggested a mixed method of sampling that addressed their concerns. Fixed-random mixture sampling schedules observations after a period of random duration after the last observation, and eliminates the fixed delay directly after an observation. The degree of randomisation required (e.g., to reduce reactivity of workers) can be selected, and the number of observations and study duration can be calculated. Liao and Pape (1996) provided the following example. The results of a preliminary study indicate that a behavioural category occurs for approximately .1 of the time and each instance of this category is, on average, 18 min in duration. A mixture of 20% fixed-interval and 80% random sampling is selected, and for a sample with 90% efficiency, the value 14.2 is taken from Table 3. $\gamma$ represents a number between 0 and 1, whereby 0 represents purely random sampling and 1 represents fixed-interval sampling. $\gamma$ is used to select a value of $d$ from Table 3 for use in Equation 7 to determine the absolute spacing between observations:

$$\text{absolute spacing} = d \cdot m \cdot (1 - p)$$ (7)
In Equation 7, $m$ is the average duration of the behavioural category, $p$ is the proportion of time of the occurrence of the category (expressed as a decimal), and $d$ is the relative observation spacing as determined from Table 3.

**Table 3.**

*Table of $d$ Values (Relative Spacing) Determined under each Efficiency Value Across a Range of Sampling Methods (from Purely Random to Purely Systematic).*

<table>
<thead>
<tr>
<th>Sampling</th>
<th>$\gamma$</th>
<th>Relative spacing ($d$)</th>
<th>Efficiency .8</th>
<th>Efficiency .9</th>
<th>Efficiency .95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>.0</td>
<td>8.0</td>
<td>18.0</td>
<td>38.0</td>
<td></td>
</tr>
<tr>
<td>10% Fixed</td>
<td>.1</td>
<td>7.1</td>
<td>16.2</td>
<td>34.2</td>
<td></td>
</tr>
<tr>
<td>20% Fixed</td>
<td>.2</td>
<td>6.3</td>
<td>14.2</td>
<td>30.2</td>
<td></td>
</tr>
<tr>
<td>30% Fixed</td>
<td>.3</td>
<td>5.4</td>
<td>12.4</td>
<td>26.4</td>
<td></td>
</tr>
<tr>
<td>40% Fixed</td>
<td>.4</td>
<td>4.6</td>
<td>10.4</td>
<td>22.4</td>
<td></td>
</tr>
<tr>
<td>50% Fixed</td>
<td>.5</td>
<td>3.9</td>
<td>806</td>
<td>18.5</td>
<td></td>
</tr>
<tr>
<td>60% Fixed</td>
<td>.6</td>
<td>3.4</td>
<td>6.8</td>
<td>14.6</td>
<td></td>
</tr>
<tr>
<td>70% Fixed</td>
<td>.7</td>
<td>3.0</td>
<td>5.2</td>
<td>10.8</td>
<td></td>
</tr>
<tr>
<td>80% Fixed</td>
<td>.8</td>
<td>2.7</td>
<td>4.2</td>
<td>7.0</td>
<td></td>
</tr>
<tr>
<td>90% Fixed</td>
<td>.9</td>
<td>2.4</td>
<td>3.4</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>Systematic</td>
<td>1.0</td>
<td>2.2</td>
<td>2.9</td>
<td>3.7</td>
<td></td>
</tr>
</tbody>
</table>


Therefore, the absolute spacing of observations in the current example would be $14.2 \times 18 \times .9 = 230$ min (Liao & Pape, 1996). When the absolute error equation is used (absolute error of $\pm 5\%$), the required number of observations is calculated to be 144 (e.g., Barnes, 1957; Niebel, 1967). However, in order for the results of the sampling to be extrapolated beyond the study period, the number of observations is divided by the efficiency value (.9).
Therefore, the new number of observations required is \( 144 / .9 = 160 \). Spacing 160 observations 230 min apart would take approximately 80 days to complete, assuming that one day is 8-hr long. Therefore only two observations could be conducted per day. If this is not desirable, the value of \( \gamma \) can be adjusted to determine a new relative spacing, a new absolute spacing, a new efficiency, and therefore a new number of observations. Liao and Pape concluded that their theoretical analysis provided an alternative to choosing either random or fixed-interval sampling, and although their analysis was based on rough estimates, fixed-interval sampling was shown to be more efficient than random sampling.

Some studies have compared continuous observation to random and fixed sampling. Knickman et al. (1992) conducted continuous observations over three or four days for each of eight medical residents. Data were collected by timing the duration of categories of resident behaviour. Finkler et al. (1993) used these data as ‘true values’ of the durations of the categories, and applied fixed-time sampling methods to determine how closely the sampled data approximated the true values. Samples were extracted from the data every 60, 30, and 15 min. The results of their analysis suggested that the samples taken every 15 min most closely approximated the continuous data (e.g., standard deviation decreased in increasing frequency of samples). However, the 15 min samples still differed greatly from the continuous data (by more than .2). Finkler et al. concluded that work sampling may not produce data representative of the full time-of-interest because the sampled data did not approximate the continuous (‘true values’) data. However, a large confound in this study was that both the number and distribution of samples were manipulated; more samples were taken as the frequency of sampling increased. Therefore, it may be the same number of samples distributed across a longer period of time yields results that more closely approximate true values than fewer but more closely distributed samples. Assumedly, this would depend on the nature of the work being studied and whether categories of behaviour occurred randomly,
cyclically, or uniformly. Randomly distributed behaviour refers to behaviour for which patterns across time cannot be identified. Behaviour that occurs cyclically may be defined as behaviour that occurs in the same repeated sequence of behaviours. For example, in a factory, the behaviour of sealing the packaging will always follow counting the number of screws into the box, which in turn will always follow retrieving the finished screws from the machine. Uniformly distributed behaviour is behaviour that occurs at the same time each hour, day, month etc. For example, in an office setting, lunch may always be taken between 12 p.m. and 1 p.m..

Wirth, Kahn, and Perkoff (1977) observed the work of physicians using both random sampling and continuous observation. Although they reported that there was no significant difference between the data collected by the two methods for 82% of the participants, each dataset was obtained at a different time. It could be argued that the data were not comparable. In addition, 26 hr of continuous observation and approximately 91 hr of random sampling were conducted (1,100 observations at an average rate of 12 per hr). Therefore, the duration of each bout of data collection was not kept constant. Although Wirth et al. concluded that random sampling is efficient and produces accurate results, true representativeness was not analysed. Analysing representativeness would have involved comparing the two sampling methods using the same dataset. It should be noted that in the medical work sampling literature, continuous observation is sometimes referred to as time-and-motion study, and random sampling referred to as work sampling (e.g., they are discussed as separate methods rather than different methods of selecting when to make observations). In the current review, continuous observation of samples shorter than the full time-of-interest, and random sampling are considered to be different ways of scheduling observations because neither consists of recording data for the full time-of-interest. In addition, observations can be scheduled to occur every second, resulting in continuous observation.
In the hypothetical example study of nurses, the work sampling expert selected a fixed-interval schedule for observations (every 5 min) due to ease of implementation (relative to using random number tables to select times). The behavioural categories under investigation were assumed to be partially cyclic in nature (e.g., a nurse was likely to meet with the next patient in the queue, take patient details, complete an assessment, make the relevant referral, and select the next patient). However, some behavioural categories were assumed to be more random. For example, a surgical resident may have approached a nurse to ask a question regarding a patient, or breaks may have been allocated based on how busy the emergency department was. Therefore, fixed-interval sampling was practical and any bias in the results was assumed to be negligible.

DATA ANALYSIS

The data collected are analysed daily so that control charts may be plotted during the study. Modern work sampling studies conducted in work settings have presented the data analyses in tables (e.g., Linden & English, 1994) or graphs (e.g., Hagerty, Chang, & Spengler, 1985). The analysis used and data displays selected are likely to reflect the purpose of the study. Therefore, the analysis conducted each day of a work sampling study is the calculation of $p$ for each category (the proportion of time spent engaged in the behaviour) and the new number of observations required by the work sampling equations (Pape, 1988). Control charts can be produced to show overall proportions of $p$, as well as the daily variation in $p$. Robinson (2010b) suggested that data can be analysed at both an individual (inter-person and intra-person variability) and group level. At the conclusion of the study, a final precision check is conducted using a control chart, and the final average $p$ value (Allderige, 1954). The average value of $p$ and total number of observations conducted are placed back into the work sampling equation (e.g., either Equation 3 or Equation 4) and the accuracy
calculated. If the final accuracy computed is less than desired, further observations are required (Jackson, 1972). Presenting the data collected by work sampling may involve combining categories into general categories (e.g., ‘productive’ and ‘unproductive’), calculating and reporting the cost (in dollars of a particular component of a process), displaying results on a graph showing changes across time, or tabulating results to show the proportions of time spent in each category (Heiland & Richardson, 1957). The results of work sampling can then be used to make changes in the work environment such as establishing time standards (Barnes, 1957), reducing undesirable work categories (Heiland & Richardson, 1957), and assessing changes in work behaviour resulting from the introduction of incentive schemes (Currie, 1963).

More recently, researchers have discussed the different statistical analyses that can be conducted with work sampling data. Miller et al. (1996) conducted analyses of independence on consecutive observations (those that occur one after another) to test the assumption that observations were independent (had no effect on one another). The analysis used by Miller et al. was a population-averaged approach rather than analysis at the level of individual workers, the results of which suggested that the variability of observations was more influenced by the dependent relationships between observations within a session than between observations across sessions. For example, the probability of observing a behaviour in two consecutive observations may be affected by the average duration of the behaviour and thus the two observations are not independent. Miller et al. suggested that this dependent relationship within a session was more likely to affect the results that the effects of conducting an observation at the same time each day. It may be that consecutive observations are not independent as assumed and therefore the variances in summary data may be larger than calculated. Tryfos (1988) also suggested that it is inaccurate to assume observations are independent, and an alternative, more complex method of determining the
required sample size may be required. Other complex statistical analyses that can be conducted on work sampling data include used variance estimates to determine the accuracy of the data (Pape, 1982). Many studies that have approached work sampling from a statistical or mathematical perspective are complex, technical, and often do not demonstrate the utility of the methods with observed or obtained data (e.g., Pape & Huang, 1991; Pape, 1992).

In the hypothetical work sampling example, the work sampling expert presented the results of the study to the management team in the form of percentages of time nurses spent engaging in each category. Control charts were also used to identify days in which the proportions of time spent in a category differed from other days. The work sampling expert assisted management in identifying factors such as short staffing or number of patients presenting to the emergency department to account for variations in the proportions of time spent in different behavioural categories.

ADVANTAGES AND LIMITATIONS OF WORK SAMPLING

Work sampling is considered to be an efficient way to measure how people spend their time in work settings. First, it permits a small number of observers to collect data on a range of people and behaviours. Work sampling is not as labour-intensive as long periods of direct, continuous observation, may be less costly that other measurement techniques, and data collection can be interrupted with little consequence (Sittig, 1993). Second, the period over which work sampling is conducted can be tailored to suit the work setting and cycles of processes (Sittig, 1993). Statistical analyses can be conducted to guide the number of observations required and the overall accuracy of the results of the study. Third, the results produced are quantifiable, minimise bias if gathered through direct observation, and the approach of work sampling is considered to be pragmatic (Oddone, Guarisco, & Simel,
These advantages have made work sampling an attractive approach to analysing work processes to a number of fields and industries.

Work sampling, however, has a number of limitations, many of which have not been addressed or explored empirically. First, although the statistical basis of the work sampling equation has been explained and manipulated (e.g., Pape, 1982), there do not appear to be any direct analyses of whether the data collected by work sampling techniques reflects true or full time-of-interest values (e.g., whether data are representative or not). For example, Abdellah and Levine (1954) suggested that momentary observations may ‘mask’ wider activities, suggesting that snap observations may not provide data indicative of how people are spending their time. Work sampling is designed to make inferences about the durations of activities, but is unable to provide information on other dimensions of behaviour, such as frequency or quality (Sittig, 1993).

Second, although methods of checking the reliability of the data obtained from different observers are sometimes attempted, it does not appear to be standard for studies to address agreement between observers. Some authors have suggested that poor inter-observer agreement invalidates results (Pelletier & Duffield, 2003), without any attempts to identify ways to increase agreement or suggest alternative methods of checking the quality (e.g., reliability, accuracy, or representativeness) of the data. Although work sampling equations can be used to check the accuracy of the data collected and calculate the number of observations required (Jackson, 1972), this is done by placing obtained values back into the equations; no attempts to compare obtained data with actual proportions of time are suggested. Another example of an indirect test of the reliability of the data is to compare work sampling data to other data collected. For example, Handa and Abdalla (1989) found that work sampling percentages correlated with unit rate productivities. Sittig (1993) suggested that relatively similar control charts between observers, in addition to a sufficiently
large sample to obtain the desired accuracy, may be sufficient. Disagreement between
observers is not directly addressed, however, nor does it objectively analyse the quality of the
data beyond the original statistics.

Third, it appears that some of the simple assumptions in work sampling, such as
reactivity to the presence of observers, are claimed with little data to suggest that this is the
case for all situations. For example, Wirth et al. (1977) illustrated reactivity when one
participant explicitly stated that they had changed their behaviour, but no attempts were made
to quantify or analyse this. Likewise, Burke et al. (2000) suggested that the bias produced in
data from workers changing their behaviour under observation diminishes over time, but
there are no studies that have attempted to explore this any further than making anecdotal
remarks.

Generally, despite a number of introductory texts written in the 1950s and 1960s that
describe work sampling (e.g., Barnes, 1957; Niebel, 1967), the methods of work sampling in
modern studies are often inconsistent and poorly described. Robinson (2010b) identified that
few studies have addressed the differences in procedures and issues in measurement or
analysis. In addition, the small number of studies that have challenged and attempted to
rectify the methods used in work sampling (e.g., Pape, 1993; Buck, Askin, & Tanchoco,
1983) are highly technical, complex, and difficult to interpret for those without advanced
knowledge of statistics (Robinson, 2010b). None of these studies have applied the theoretical
analyses to real data, nor do the studies attempt to examine the utility of these analyses in
work settings. Roll and Yadin (1986) claimed that work sampling practitioners have obtained
poor results under traditional work sampling, but did not describe the limitations of the data
obtained by these practitioners nor provided a reference for this claim. Roll and Yadin
suggested that some data limitations may have been due to erroneous assumptions
underpinning the methods of work sampling (e.g., regarding the distributions of the
behaviours of interest). As an alternative model, Roll and Yadin suggested the use of a multi-parameter distribution, in which the proportion of time spent in an activity is determined by a number of quantifiable variables such as the spacing and timing of observations. However, they did not provide any empirical evidence for their claim that their suggested change in approach could improve work sampling methods.

Zheng, Guo, and Hanauer (2011) conducted a review of time-and-motion studies using continuous direct observation methods over a twenty year period. They found that the methods used differed greatly and were not described in enough detail for replication. Zheng et al. found that they were also forced to exclude studies that, despite the use of the term time-and-motion, did not conduct a time-and-motion study as they had defined it, further highlighting the confusion in terminology and methods. As a solution, Zheng et al. developed a flow chart to assist researchers to select work sampling methods based on the setting in which the study is to be conducted, the observers used, the type of person being observed, the desired research design, and the type of data recording and analysis.

Although many studies provide a logical or mathematical rationale for the use of the work sampling equations, the equations do not appear to have been thoroughly researched empirically. In addition, more modern studies conducted in the workplace do not appear to use the work sampling equations. For example, Mathiassen, Burdorf, van der Beek, and Hansson (2003) compared the accuracy of the data on job exposure estimates obtained using several sampling methods. Their data comparing sampling methods for measuring upper trapezius activity in workers was used to develop a decision tree. Practitioners are able to use the decision tree to choose a sampling method most appropriate for their job exposure data collection. Although Mathiassen et al. (2003) were concerned with measuring representatively the types of behaviours that have been measured using the work sampling equation, and their method of data collection (e.g., random and fixed-interval sampling) was
similar to that of work sampling, no equation was used. Their study is one of many modern ergonomics studies concerned with representative data collection. It appears that work sampling equations are no longer widely used, despite a lack of research demonstrating the utility (or otherwise) of the equations.

A potential reason for the absence of work sampling equations in modern studies could be a shift in focus to other factors outside sampling method that influence the representativeness of samples. For example, when selecting people from whom to obtain information regarding job components (which may form the basis of behavioural categories to be measured through direct observation), the choice of informant can affect the components selected. Informants may be influenced by factors such as being pressured to conform to others’ verbal reports of the job components, attempts appease or impress management, or an inability to articulate job components accurately (Morgeson & Campion, 1997). Although specific to the methods of job analysis, the modern literature reflects an interest in other factors affecting representativeness such as those described by Morgeson and Campion (1997), and contains discussions regarding distinguishing accuracy from validity and agreement (e.g., Harvey & Wilson, 2000). The modern literature illustrates a shift in focus away from work sampling methods more popular in the 1950s and 1960s (i.e., the equations).

Robinson (2010b) identified other factors that may have influenced the shift away from work sampling equations. For example, some work tasks involve private events such as analysing data or writing reports whose directly observable behaviours (e.g., in writing, motor movements of the hand) may give little information about the content. Alternatively, permanent products (i.e., in the present example, what was written) could be used to assess the content. Although self-report could be used to capture the details regarding difficult-to-observe tasks, self-report can be inaccurate. In addition, although broad (and therefore
higher-duration) categories of behaviour such as routine and non-routine work may be
directly observed easily (e.g., Gold, Park, & Punnett, 2006), lower-duration behaviours may
not be as easily captured representatively. Lobb and Woods (2012) showed that low-duration
construction tasks were not sampled representatively using fixed-interval sampling. Low-
duration behaviours will require a large number of observations when the relative accuracy
work sampling equation (Equation 3) is used, or a very small number of observations when
the absolute accuracy work sampling equation (Equation 4) is used. Large numbers of
observations are unlikely to be practical, and small numbers of observations are unlikely to
produce representative data. Therefore, work sampling using the equations may not be used
in more modern studies when the behaviours of interest are of lower-duration or more
variable, like those measured by Lobb and Woods (2012).

Perhaps it is the adoption of work sampling methods across a diverse range of fields
that has contributed to the current confusion and lack of empirical analysis of the methods.
Some aspects of work sampling that may contribute to the lack of analyses of the method
may include the use of the equation with small $p$ values and the subsequent large number of
required observations, empirically-derived methods of estimating the initial $p$ value to be
placed into the equation, a comparison between the relative and absolute accuracy equations.
Perhaps most importantly, research is required to determine whether work sampling does
produce data indicative of overall proportions of time spent in different activities.

EXAMPLES OF WORK SAMPLING STUDIES

A review of 44 work sampling studies from the period 1954 to 2011 was conducted to
obtain information on the purpose and methods of work sampling studies (Table 4). The 44
studies were a selection of work sampling studies published in peer reviewed journals, and
were selected by searches of science databases (e.g., Web of Science). Although this was not
an exhaustive review, all work sampling articles for which the full text could be obtained from the databases were included and identified through the use of the key words ‘work sampling’. Articles were discarded if the term ‘work sampling’ had been used to describe a different time-and-motion study method (e.g., the terms had been confused), or if the study involved ‘work sampling’ as used in the education sector (an unrelated method). Studies did not have to present graphical or tabulated data for inclusion, but were included if details of the method (e.g., the participants, duration of the study) were provided.

Of the studies reviewed, 27 were conducted in hospital settings (61.4%), five in other medical settings, four in nursing homes or geriatric units, two in businesses, one on a construction site, one in a private home, one in a museum, one in a restaurant, one in a factory, and one in a university. Although this is not a complete literature of all work sampling studies published, it could be suggested that the field that has most obviously adopted work sampling methodology for research is the medical field. Perhaps surprisingly, there were very few studies conducted in traditional work sampling settings such as factories. The most frequently studied population in the 44 studies was nurses (19 studies, 43.2% of the studies), followed by physicians or other medical practitioners (11 studies, 25%).

Despite the presence and discussion of the equations representing the underlying assumptions of work sampling in introductory texts such as Barnes (1957) and technical papers such as Pape (1988), very few of the studies reviewed either mentioned or used any equations. Eight studies (Ampt, Westbrook, Creswick, & Mallock, 2007; Domenech, Payton, Hill, & Shukla, 1983; Gunesoglu & Meric, 2006; Kelly, 1964; Lind & Hill, 1974; Munyisia, Yu, & Hailey, 2011; Myny et al., 2009; Robinson, 2010a) used the equation directly (18.2% of the studies). One study discussed the use of the equation but did not appear to use it (Gardner et al., 2010). The lack of use of the work sampling equations suggests a deviation from the original intention of work sampling to provide an empirical method of selecting the
number of observations for a representative sample. It could be argued, however, that because the work sampling equations do not appear to have been tested empirically, it is difficult to argue that the use of the equations is required. Of the eight studies that used work sampling equations, three used the relative accuracy equation, four used the absolute accuracy equation, and it could not be determined which was used by one study (Ampt et al., 2007). Of the studies reporting confidence intervals, all but one used 95% confidence intervals and one study did not report the confidence interval used (Gunesoglu & Meric, 2006). Ampt et al. (2007) used 90% confidence intervals. A range of ± 1% accuracy to ±10% accuracy was used across these studies. None of the studies analysed the final accuracy of the data, nor did they employ the use of control charts or any other statistical work sampling technique. Crance and Willoughby (2007) reported the use of control charts but they did not present them in the article.

The duration of the studies ranged from two days to 365 days, and the total number of observations conducted in each study ranged from 111 to 101,325. The duration of the studies however, was influenced by the sampling method selected. Regarding measurements used, 14 studies employed duration or continuous measures (31.8%), 17 studies employed fixed-interval sampling (38.6%), and eleven studies employed random sampling (25%) (two studies did not report the sampling schedule). It should be noted, however, that some studies that reported random sampling actually conducted stratified non-continuous random sampling, in which observations are randomly scheduled within blocks of time (Pape, 1988). Data collection methods varied between studies, but the majority used pen-and-paper methods (84.1%). The remainder used computers, including hand-held and touch-screen computers. Seven studies used self-report methods for data collection, none of which conducted reliability checks on the data. Only 13 studies conducted reliability checks, however all conducted these checks through inter-observer agreement (IOA) measures; no
studies used any other methods to check the reliability or accuracy of the data. It should also
be noted that many of these studies conducted IOA checks during training and not during the
data collection period (e.g., Gardner et al., 2010), limiting the ability of researchers to claim
that the data collected were reliable.

All of the studies that reported the ways in which the data were analysed reported the
use of descriptive statistics (43 out of 44 studies). The analysis was usually presented in
graphs or tables. Some studies conducted statistical tests when the purpose was to evaluate
differences between groups of people or activities. One study compared the obtained
durations of activities to pre-existing standards (Bartholomewicz & Hunstein, 2004),
however the primary purpose of most studies was to determine the durations of the
behaviours of interest. For example, Lind & Hill (1974) measured duration to evaluate
staffing levels. The purpose of some studies was to evaluate the data collection method (e.g.,
Caughey & Chang, 1998), or evaluate staff performance (e.g., Finlay, Norman, Stolberg,
Weaver, & Keane, 2006).

The reviewed selection of studies demonstrated a range of methods, aims, and settings
in empirical work sampling studies. However, the lack of adherence to the underlying
principles of work sampling, and the often incomplete reporting of the method used to
conduct the study illustrates the claim made by Robinson (2010b) that work sampling is
plagued by inconsistencies. It appears that analysing the utility of work sampling methods,
assessing the reliability, accuracy, and representativeness of work sampling data, and
addressing methodological variations is of great importance should researchers and
practitioners continue to use work sampling methods.
<table>
<thead>
<tr>
<th>Author &amp; Date</th>
<th>Equation reported</th>
<th>Setting / Participants</th>
<th>Study duration</th>
<th>Number of observations</th>
<th>Sampling method</th>
<th>Observers</th>
<th>Data collection</th>
<th>Reliability Check</th>
<th>Data analysis</th>
<th>Aim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdellah &amp; Levine (1954)</td>
<td>No</td>
<td>Hospital / Nurses</td>
<td>5 days</td>
<td>Unclear</td>
<td>Fixed-interval / 5 min</td>
<td>Nurses</td>
<td>Pen-and-paper</td>
<td>No</td>
<td>Descriptive statistics</td>
<td>Identify durations of activities</td>
</tr>
<tr>
<td>Ampt, Westbrook, Creswick, &amp; Mallock (2007)</td>
<td>90% CI, 5% accuracy</td>
<td>Hospital / Nurses</td>
<td>91 days</td>
<td>3,910</td>
<td>Random intervals / 32 per hr</td>
<td>Health staff</td>
<td>Pen-and-paper</td>
<td>IOA</td>
<td>Descriptive statistics / significance tests</td>
<td>Compare self-report and direct observation</td>
</tr>
<tr>
<td>Bartholomeyczik &amp; Hunstein (2004)</td>
<td>No</td>
<td>Homes / Family caregivers</td>
<td>Unclear / n/a</td>
<td>Duration timed (onset and offset)</td>
<td>Nursing students</td>
<td>Pen-and-paper</td>
<td>No</td>
<td>Descriptive statistics / comparison to standards</td>
<td>Identify durations of activities</td>
<td></td>
</tr>
<tr>
<td>Buchholz, Paquet, Punnett, Lee, &amp; Moir (1996)</td>
<td>No</td>
<td>Construction site / Workers</td>
<td>3 days</td>
<td>Unclear</td>
<td>Fixed-intervals / 45 s</td>
<td>Unclear</td>
<td>PATH Tool</td>
<td>IOA</td>
<td>Descriptive statistics</td>
<td>Pilot study of observation tool</td>
</tr>
<tr>
<td>Burke et al. (2000)</td>
<td>No</td>
<td>Hospital / Nurses</td>
<td>25 days</td>
<td>n/a</td>
<td>Duration timed (onset and offset)</td>
<td>Self-report and researchers</td>
<td>Pen-and-paper</td>
<td>No</td>
<td>Descriptive statistics</td>
<td>Compare self-report to sampling</td>
</tr>
<tr>
<td>Cardona, Tappen, Terrill, Alcosta, &amp; Eusebe (1997)</td>
<td>No</td>
<td>Geriatric home / Nurses</td>
<td>50 days</td>
<td>2,040</td>
<td>Fixed-interval / 15 min</td>
<td>Unclear</td>
<td>Pen-and-paper</td>
<td>IOA</td>
<td>Descriptive statistics</td>
<td>Identify durations of activities</td>
</tr>
<tr>
<td>Study</td>
<td>Setting</td>
<td>Sample Size</td>
<td>Duration</td>
<td>Frequency</td>
<td>Methodology</td>
<td>Data Collection</td>
<td>Outcomes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----------------------------</td>
<td>-------------</td>
<td>----------</td>
<td>-----------</td>
<td>-------------</td>
<td>-----------------</td>
<td>-----------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caughy &amp; Chang (1998)</td>
<td>Nursing home / Staff</td>
<td>2 or 7 days</td>
<td>n/a</td>
<td></td>
<td>Self-report</td>
<td>Barcode wands</td>
<td>Onset and offset recorded</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chaboyer et al. (2008)</td>
<td>Hospital / Nurses</td>
<td>10 days</td>
<td>14,528</td>
<td></td>
<td>Research assistants</td>
<td>Pen-and-paper</td>
<td>Fixed interval 10 min</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Domenech, Payton, Hill, &amp; Shukla (1983)</td>
<td>Hospital / Physical Therapists</td>
<td>30 days</td>
<td>5,824</td>
<td></td>
<td>Researcher</td>
<td>Pen-and-paper</td>
<td>Duration / Fixed-interval 15, 30 and 60 min</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finkler, Knickman, Hendrickson, Lipkin, &amp; Thompson (1993)</td>
<td>Hospital / Residents</td>
<td>Unclear</td>
<td>1,561</td>
<td></td>
<td>Pre-medical students</td>
<td>Pen-and-paper</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finlay, Norman, Stolberg, Weaver, &amp; Keane (2006)</td>
<td>Hospital / Student radiologists</td>
<td>365 days</td>
<td>111</td>
<td></td>
<td>Training evaluator</td>
<td>Hand-held computer</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fontaine, Speedie, Abelson &amp; Wold (2000)</td>
<td>Hospital / Physicians and assistants</td>
<td>Unclear</td>
<td>1,826</td>
<td></td>
<td>Unclear</td>
<td>Pen-and-paper</td>
<td>Fixed-interval rotation / 8 min</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Design and Measurement Details</td>
<td>Activities</td>
<td>Duration</td>
<td>Methodology</td>
<td>Data Collection</td>
<td>Data Analysis</td>
<td>Notes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>------------------------------------------------------------------------------------------------</td>
<td>-------------</td>
<td>------------</td>
<td>--------------</td>
<td>----------------</td>
<td>---------------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gardner et al. (2010)</td>
<td>Discuss but not used directly</td>
<td>Hospital / Nurses</td>
<td>Unclear</td>
<td>12,189</td>
<td>Fixed-interval / 10 min</td>
<td>Unclear</td>
<td>Pen-and-paper</td>
<td>IOA</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gunesoglu &amp; Meric (2006)</td>
<td>Relative formula, 1% accuracy</td>
<td>Factory / Sewing room workers</td>
<td>16 days (2 per week)</td>
<td>14,000</td>
<td>Random sampling rotation</td>
<td>Unclear</td>
<td>Pen-and-paper</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hagerty, Chang, &amp; Spengler (1985)</td>
<td>No</td>
<td>Hospital / Nurses</td>
<td>14 days</td>
<td>n/a</td>
<td>Duration estimated (min)</td>
<td>Self-report</td>
<td>Pen-and-paper</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hoffman, Tasota, Scharfenberg, Zullo, &amp; Donahoe (2003)</td>
<td>No</td>
<td>Medical centre / Nurses and physicians</td>
<td>147 days</td>
<td>Unclear</td>
<td>Fixed-interval / 10 and 5 min</td>
<td>Nursing students</td>
<td>Clinician Activities Tool</td>
<td>IOA</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Significance tests between groups</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hollingsworth, Chisholm, Cordell, &amp; Nelson (1998)</td>
<td>No</td>
<td>Hospital / Physicians and nurses</td>
<td>40 days</td>
<td>Unclear – 180-min sessions</td>
<td>Duration timed (onset and offset)</td>
<td>Unclear</td>
<td>Pen-and-paper</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Compare acute care nurse and physicians</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kelly (1964)</td>
<td>Absolute error, 95% CI, 2% accuracy</td>
<td>Businesses /Executives</td>
<td>21 days</td>
<td>2,800</td>
<td>Random sampling</td>
<td>Researcher</td>
<td>Pen-and-paper</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Setting</td>
<td>Duration</td>
<td>Methodology</td>
<td>Data Collection</td>
<td>Findings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>----------------------------------</td>
<td>----------------</td>
<td>-------------</td>
<td>----------------</td>
<td>-----------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knickman, Lipkin, Finkler, Thompson, &amp; Kiel (1992)</td>
<td>Hospital / Physicians-in-training</td>
<td>7 days n/a</td>
<td>Duration timed (onset and offset)</td>
<td>Pre-medical students</td>
<td>Pen-and-paper No</td>
<td>Descriptive statistics Reallocate physician work to non-physicians</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linden &amp; English (1994)</td>
<td>Hospital / Nurses</td>
<td>14 days 6,709</td>
<td>Fixed-interval 12 min / duration</td>
<td>Unclear Computer No</td>
<td>Descriptive statistics Compare cost and quality of work</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lind &amp; Hill (1974)</td>
<td>University / Office staff</td>
<td>28 days 246</td>
<td>Random sampling</td>
<td>Unclear Pen-and-paper No</td>
<td>Descriptive statistics Evaluate staffing levels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lo et al. (2007)</td>
<td>Hospital / Physicians</td>
<td>Unclear 30 (unclear duration)</td>
<td>Duration timed (onset and offset)</td>
<td>Research assistants Computer No</td>
<td>Descriptive statistics Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mavis, Pearson, Stewart, &amp; Keefe (2009)</td>
<td>School health clinic / Clinic staff</td>
<td>6 days 1,492</td>
<td>Fixed-interval / 15 min</td>
<td>Self-report Pen-and-paper No</td>
<td>Descriptive statistics Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mayer (1992)</td>
<td>Hospital / Nurses</td>
<td>Unclear</td>
<td>Fixed-intervals / 5 min</td>
<td>Unclear Pen-and-paper No</td>
<td>Descriptive statistics Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Setting Type</td>
<td>Setting Details</td>
<td>Duration (hrs)</td>
<td>Duration Method (onset and offset)</td>
<td>Data Collection</td>
<td>Findings</td>
<td>Research Methods</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>--------------</td>
<td>-----------------</td>
<td>----------------</td>
<td>----------------------------------</td>
<td>----------------</td>
<td>----------</td>
<td>------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Milosavljevic, Williams, Perez, &amp; Dalla (2011)</td>
<td>No Hospital</td>
<td>Unclear</td>
<td>287</td>
<td>Duration timed</td>
<td>Dietetics</td>
<td>No</td>
<td>Descriptive statistics Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morton et al. (2004)</td>
<td>No Hospitals</td>
<td>6 days</td>
<td>8,640</td>
<td>Duration timed</td>
<td>Students</td>
<td>No</td>
<td>Descriptive statistics Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Munyisia, Yu, &amp; Hailey (2011)</td>
<td>No Nursing</td>
<td>5 days</td>
<td>5,444</td>
<td>Fixed-interval rotation / 9 or 5 min</td>
<td>Researcher</td>
<td>No</td>
<td>Descriptive statistics Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myny et al. (2009)</td>
<td>No Hospital</td>
<td>40 days</td>
<td>13,292</td>
<td>Random sampling</td>
<td>Nurses</td>
<td>IOA</td>
<td>Descriptive statistics Determine standard time per activity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norbergh, Asplund, Rassmussen, Nordahl, &amp; Sandman (2001)</td>
<td>No Geriatric unit</td>
<td>4 days</td>
<td>2,024</td>
<td>Fixed-interval / 10 min</td>
<td>Nurses</td>
<td>IOA</td>
<td>Descriptive statistics Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oddone, Guarisco, &amp; Simel (1993)</td>
<td>No Hospital</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Random intervals / average 3.2 per hr</td>
<td>Self-report</td>
<td>No</td>
<td>Descriptive statistics Compare estimates with self-report</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study (Year)</td>
<td>Design</td>
<td>Setting</td>
<td>Sample Size</td>
<td>Duration</td>
<td>Data Collection Method</td>
<td>Data Analysis</td>
<td>Purpose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>--------</td>
<td>--------------------------------------</td>
<td>-------------</td>
<td>----------</td>
<td>------------------------</td>
<td>---------------</td>
<td>-------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quach et al. (2011)</td>
<td>No</td>
<td>Clinic / Vaccine staff</td>
<td>3 days per site</td>
<td>n/a</td>
<td>Duration timed (onset and offset)</td>
<td>Unclear</td>
<td>Compare duration of two activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quist (1992)</td>
<td>No</td>
<td>Hospital / Nurses</td>
<td>84 days</td>
<td>Unclear</td>
<td>Fixed-interval / 10 or 15 min</td>
<td>Nurses</td>
<td>Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Robinson (2010a)</td>
<td>Relative formula, 95% CI, 10% accuracy</td>
<td>Company / Engineers</td>
<td>20 days</td>
<td>1,000</td>
<td>Stratified, non-contin. sampling</td>
<td>Self-report</td>
<td>Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stevenson, Caverly, Srebnik, &amp; Hendryx (1999)</td>
<td>No</td>
<td>Mental health centre / Staff</td>
<td>70 days</td>
<td>1,033</td>
<td>Random intervals / 2 per day</td>
<td>Self-report</td>
<td>Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turnbull, MacFadyen, van Barneveld, &amp; Norman (2000)</td>
<td>No</td>
<td>Hospital / Student clinicians</td>
<td>56 days</td>
<td>Average of 19 per 67 students</td>
<td>Unclear</td>
<td>Supervisor Pen-and-paper</td>
<td>Descriptive statistics Test feasibility of tool</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upenieks (1998)</td>
<td>No</td>
<td>Hospital / Nurses</td>
<td>10 days</td>
<td>2,835</td>
<td>Fixed-interval rotation / 20 min</td>
<td>Unclear</td>
<td>Pen-and-paper IOA</td>
<td>Descriptive statistics Improve work flow</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urden &amp; Rood (1997)</td>
<td>No</td>
<td>Hospital / Nurses</td>
<td>42 days</td>
<td>101,325</td>
<td>Fixed-interval / 10 min</td>
<td>Managers Pen-and-paper IOA</td>
<td>Descriptive statistics Identify durations of activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Setting</td>
<td>Sample</td>
<td>Duration</td>
<td>Data Collection</td>
<td>Data Analysis</td>
<td>Objective</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>--------</td>
<td>----------</td>
<td>----------------</td>
<td>--------------</td>
<td>-----------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weigel, Muller, Zupanc, &amp; Angerer (2009)</td>
<td>Hospital / Physicians</td>
<td>35 days</td>
<td>n/a</td>
<td>Onset and offset</td>
<td>Clinician / Student</td>
<td>Pen-and-paper</td>
<td>IOA</td>
<td>Descriptive statistics / significance tests</td>
<td>Test feasibility of categories</td>
<td></td>
</tr>
<tr>
<td>Williams, Harris, &amp; Turner-Stokes (2009)</td>
<td>Neuro. unit / Nurses</td>
<td>14 days</td>
<td>8,883</td>
<td>Fixed-interval rotation / 5 min</td>
<td>Unclear</td>
<td>Pen-and-paper</td>
<td>IOA but not of same data</td>
<td>Descriptive statistics</td>
<td>Identify durations of activities</td>
<td></td>
</tr>
<tr>
<td>Wirth &amp; Kahn (1977)</td>
<td>Hospital / Physicians</td>
<td>28 days</td>
<td>1,100</td>
<td>Duration / random sampling 12 per hr</td>
<td>Unclear</td>
<td>Pen-and-paper</td>
<td>Accuracy checked by supervisor</td>
<td>Descriptive statistics</td>
<td>Compare continuous observation to sampling</td>
<td></td>
</tr>
</tbody>
</table>
WORK SAMPLING AND APPLIED BEHAVIOUR ANALYSIS: RATIONALE

The adoption of work sampling methods across a range of settings, industries, and behaviours suggests that work sampling may offer a method of efficiently collecting representative data on the durations of activities in which people engage. Applied behaviour analysis (ABA) would benefit from the introduction of a method of choosing a sampling method that permits practitioners and researchers to collect representative data efficiently (i.e., in a shorter period of time than direct observation of the full time-of-interest without a loss of the quality of the data). Although general rules (i.e., "this is better than that") can be derived currently from the behaviour analytic literature, work sampling may provide a data-based method for selecting sampling methods. The utility of work sampling concepts, equations, and methods in ABA is thus an area of interest and potential study. What follows is a comparison of some of the common aspects of work sampling and behaviour analysis.

There are several aspects of ABA that are analogous to aspects of work sampling and the similarities between the two support the notion that there may be some utility in work sampling for behaviour analysis. Behaviour analysis can be concerned with the durations of behaviours or how people spend their time. Altmann (1974) indicated that duration measures are used in behaviour analysis when the behaviour of interest is a behavioural state, rather than a discrete event. For example, Schumacher and Rapp (2011) used duration measures to record the duration of compliant sitting of child with autism with the aim to increase the duration of sitting in order for the child to have a haircut. Himadi and Curran (1995) used
duration recording to measure the duration of auditory hallucinations in a woman diagnosed with mental illness with an aim to decrease the duration of the hallucinations.

**Direct observation.** Both work sampling and behaviour analysis use observable measures rather than subjective or covert measures. Similar to the poor correspondence between indirect measures such as self-report and true values reported in work sampling (Ampt, Westbrook, Creswick, & Mallock, 2007), behaviour analytic researchers have suggested similar poor correspondence between indirect measures and true values in their field (e.g., Bellack & Hersen, 1977). It should be noted, however, that neither behaviour analysis nor work sampling reject self-report measures entirely; direct measures are preferred (Kahng, Ingvarsson, Quigg, Seckinger, & Teichman, 2011). Discontinuous work sampling uses what is akin to momentary time sampling (MTS) in behaviour analysis, with different schedules of MTS within an observation session (e.g., random, fixed, and non-continuous). Work sampling might also be done in continuous observations; the daily observations conducted in a continuous block (e.g., across consecutive seconds) as a manager may do in a factory. Similarly, a behaviour analytic consultant may conduct an continuous observation in a classroom upon receiving a referral to assess a child’s challenging behaviour.

Work sampling studies are concerned with the number of observations to be conducted (sample size), as specified by the work sampling equations (e.g., Morris, 1969). In behaviour analysis, the issue of sample size may be conceived as the issue of how much data to collect, a question often answered through visual analysis of the data collected. For example, a behaviour acquisition programme may be terminated upon attainment of a mastery criterion (e.g., Luciano, Herruzo, & Barnes-Holmes, 2001). Another example may be choosing how many baseline data points to collect, a decision influenced by the variability in the baseline (e.g., visual inspection of the data), the severity of the behaviour, and constraints of the setting.
MTS can be used to observe multiple people in a rotation or multiple behaviours (Adams, 1991). For example, Hanley, Cammilleri, Tiger, and Ingvarsson, (2007) used MTS to record the behaviours of 20 children per observation session. Observing multiple participants or behaviours can increase the efficiency of the sampling method. MTS has been used in behaviour analysis to record mutually exclusive categories of behaviour in coding systems similar to some systems reported in work sampling. For example, McIver, Brown, Pfeiffer, Dowda, and Pate (2009) developed and tested the Observational System for Recording Physical Activity in Children-Home (OSRAC-H). The OSRAC-H coding system contained mutually exclusive categories, recorded through 5-s MTS observations conducted every 25-s, and was similar to examples in work sampling such as the PATH tool (Buchholz, Paquet, Punnett, Lee, & Moir, 1996). Both work sampling and behaviour analysis emphasize the importance of thorough and clear definitions of the behaviour to be recorded.

Settings. The settings in which work sampling and behavioural studies are conducted can be similar. Although the field of time-and-motion studies developed in settings in which the work done is typically repetitive or cyclic (e.g., factories), work sampling is recommended for use in settings in which behaviours are not cyclic (Mundel, 1960). However, the claim that work sampling is best suited to such settings does not appear to have been tested empirically. Although not exclusively, behaviour analysis is also often conducted in settings in which behaviour is not cyclic. Work sampling often measures the behaviour of workers or staff, a population that is also observed in behaviour analysis. For example, Norbergh, Aspslund, Rassmussen, Nordahl, and Sandman (2001) conducted a work sampling study on the durations of activities in which patients engaged in a geriatric unit, and Mansell and Beadle-Brown (2011) measured staff interaction with residents in residential institutions. Despite the common measurement of the behaviours of workers, work sampling methods could be used to observe any individual's behaviour when the duration of the behaviour is of
interest. It should be noted, however, that work sampling typically evaluates group data, whereas behaviour analysis is most often concerned with individual behaviour.

**Data analysis.** Both work sampling and behaviour analysis use visual analyses of the data (e.g., graphs) to present results and make decisions about environmental changes. Although visual inspection of graphs can be subjective in behaviour analysis, it should permit a behaviour analyst to predict the level, trend, and stability of the data (Roane, Ringdahl, Kelley, & Glover, 2011). Visual inspection of graphs in work sampling is aided by the use of control charts and limits upon which decisions can be made (Heiland & Richardson, 1957). Data analysis in behaviour analysis typically describes changes in dependent variables as independent variables are manipulated, but more in-depth analyses such as calculation of conditional probabilities may be conducted (e.g., Thompson & Borrello, 2011).

**Statistical Process Control.** Statistical Process Control (SPC), from which the control charts used in work sampling are derived, has also been suggested for use in behaviour analysis. Mawhinney (1992) suggested that behaviour analysis and SPC are both concerned with variability in what is being measured. The data collected through sampling in behaviour analysis can be graphed in a control chart similarly to data collected through analysing a process using work sampling, and data points that fall outside control limits can be analysed. Visual analysis of graphed data is widely used in behaviour analysis (Hagopian et al., 1997) and this is the method of analysis of control charts. Mawhinney (1992) suggested that control charts could be used to quantify customer satisfaction through the rejection of unacceptable outlying data points (with subsequent changes to the environment to prevent any further outliers), or by minimising the variables likely to cause outliers. Customer satisfaction is the concept of social validity in behaviour analysis. Therefore, SPC can be used to determine the effectiveness of a behavioural programme through analysing variability. Pfadt and Wheeler (1995) suggested that SPC is a problem-solving, easy-to-use
tool that transforms behavioural data into attribute data like the data collected in work sampling.

Pfadt, Cohen, Sudhalter, Romanczyk, and Wheeler (1992) suggested that SPC could be used in behaviour analysis to evaluate baseline stability. Stability in baseline measurements will permit a researcher to evaluate the effect of independent variables on dependent measures. Pfadt et al. evaluated previously collected baseline data by re-plotting the data on a control chart. They demonstrated that the control chart allowed for evaluation of variability that may not have been detected through visual inspection of the original graph. It could be argued that the purpose of the behavioural intervention would dictate whether the behaviour analyst discarded these data as outside control; data points outside the lower control limit may show unimportant variability if the purpose of the behavioural intervention is to teach a functional skill. Pfadt and Wheeler (1995) suggested using the SPC criteria for stability suggested by Wheeler and Chambers (1992) to determine whether a baseline is stable. In addition, the identification of data points outside upper control limits (e.g., ‘bad days’ for challenging behaviour) may encourage the behaviour analyst to identify variables occurring on these days that may be responsible for the increase in the behaviour (e.g., the presence of particular staff) (Pfadt & Wheeler, 1995). Redmon (1992) suggested that just as SPC may be of use in applied behaviour analysis, behaviour analysis may assist in measurement methods used in SPC. This may be true of work sampling also, because the behaviour analytic literature has much to offer with regard to sampling methods.

Reliability. Interobserver agreement (IOA) has been reported in modern work sampling studies (e.g., Ampt et al., 2007). IOA is calculated to determine the extent of agreement between two independent observers (Bailey & Burch, 2002) and indicates the reliability of the data. Although IOA does not provide a measure of accuracy (as both observers may be inaccurate but produce high agreement), it assists in choosing to make
changes to observation systems and behavioural definitions (Kahng et al., 2011). Many work sampling studies calculate IOA during observer training as a training criterion, but do not evaluate IOA during the study (e.g., Gardner et al., 2010). In contrast, the behaviour analytic literature suggests that IOA should be conducted in 25% to 30% of observations across all conditions (e.g., baseline and intervention) (Bailey & Burch, 2002). IOA is a commonly reported method of checking data in work sampling studies, but does not provide a measure of representativeness.

**Accuracy.** When low-duration behaviours are of interest in work sampling studies, more lenient criteria for accuracy may be used. For example, Barsness and Trinca (1978) used ± 10% relative accuracy when measuring pharmacist behaviours occurring for 15% of the time-of-interest. Also, many work sampling studies have been concerned with higher-duration behaviours. For example, Myny et al. (2009) measured categories such as direct and indirect patient care, which accounted for 25.8% and 21.8% of nurses’ time respectively.

However, low-duration behaviours are often of interest in applied behaviour analysis, and there are some low-duration behaviours for which lenient criteria for accuracy is not acceptable. For example, low-duration behaviour such as self-injurious eye-gouging may be required to be measured within very stringent accuracy limits. When low-duration behaviours are of interest, the relative accuracy work sampling equation (Equation 3) requires a large number of observations, and the absolute accuracy work sampling equation (Equation 4) requires a small number of observations. Conducting large numbers of observations may be impractical in settings in which behaviour analysts may work, however Equation 4 is unlikely to be of use if low-duration behaviours may be missed. Therefore, the degree of resolution (i.e., acceptable error) may be different across work sampling in industrial and work settings when compared with behaviour analytic settings. The difference in acceptable error is likely to be related to the purpose of data collection. For example, behaviour analysts
may require stringent accuracy limits due to ensure that data on which assessments, intervention selection, and measurement of intervention effectiveness are based is representative. By contrast, less stringent accuracy limits may be acceptable in work sampling settings when measuring the components of work for a job analysis. The impact of biased data may differ across the two types of setting.

Combining work sampling and behaviour analysis. Behaviour analysis and work sampling have several aspects in common, but only one published study has appeared to develop some links between work sampling and behavioural data. Besterfield-Sacre, Shuman, Wolfe, Clark, and Yildirim (2007) found that work sampling methods produced behavioural data representative of continuous observation of the whole time-of-interest. Observers were trained initially in continuous recording, and recorded mutually exclusive categories of subjective teamwork behaviours such as ‘working together’ and ‘coming to conclusions’ from 90-min videotaped sessions. Each participant was observed individually, and observers were able to watch the tapes multiple times. Observers worked in groups of two or three until acceptable IOA was obtained. The proportions of time spent in each category were averaged across observers to determine the ‘true values’ from the continuous recording of the whole time-of-interest (p value range of 0 to .62; the behaviour occurred for between 0% and 62% of the time-of-interest). A second group of observers viewed the same videos but used either fixed interval 10-s or 20-s sampling or floating-duration interval sampling to record the proportions of time of each behavioural category. Besterfield-Sacre et al. used the relative accuracy work sampling equation (Equation 3) to determine the number of observations to be conducted in the floating-duration interval sampling with confidence level of 95% and relative accuracy of ± 20%. As the desired accuracy was not a very stringent value, the number of observations calculated was small. Besterfield-Sacre et al. reported calculating 45 required observations, but the relative equation as reported in texts
such as Barnes (1957) calculates 100 required observations. The version of the equation used by Besterfield-Sacre et al. appeared to differ from that reported by Barnes, illustrating the inconsistencies in the reported methods of work sampling. Floating-duration interval sampling was conducted similarly to the non-continuous random sampling reported by Pape (1988), in which the 90-min video was divided by 45 and observations scheduled at a random time within each 2-min interval. Besterfield-Sacre et al. found that floating-duration interval sampling produced the most representative data, followed by 10-s interval and 20-s interval sampling respectively (i.e., shorter intervals produced more representative results). However, the study was conducted over a relatively short period (90 min), with no independent IOA, and with somewhat subjective behavioural categories. Therefore, work sampling and behaviour analysis could be combined in a study that evaluates a longer time-of-interest, a range of \( p \) values across behavioural categories, the utility of the different work sampling equations, and of different sampling methods.

**Rationale for Study 1.** The potential benefit of the adoption of work sampling methods in behaviour analysis is the ability to select the number of MTS observations required to obtain a representative sample of a behaviour of a particular duration. A data-based method of choosing the number of MTS observations is preferable to choosing the number of MTS observations based on more general findings that shorter intervals produce more representative samples (e.g., Brulle & Repp, 1984).

Study 1 was an exploratory analysis of the use of the work sampling equations for behaviours of varied duration in a setting in which applied behaviour analysis is typically used (i.e., whether the equation identified the number of observations required for a particular level of accuracy). Following preliminary observations, the work sampling equations were used to select the number of required observations, which were then extracted as discontinuous MTS (fixed-interval, random, or non-continuous) or continuous observation
sessions. Continuous observation sessions were analogous to observations conducted in applied setting, such as a school classroom, during which a practitioner may observe continuously for a period of time (e.g., 60 min) that is convenient for the people in the setting and the practitioner. The timing of the start of continuous observation sessions was also varied to assess the effect of start time on the representativeness of the samples of behaviour.

To test the Rule of 1,000, the required 1,000 MTS were extracted from the full week of observation recording for each participant evenly distributed across days. As one of the work sampling equations (Equation 3 using relative accuracy) sometimes produces impractically large numbers of observations, the representativeness of MTS conducted at intervals reported to be practical in classroom settings (i.e., 5-min MTS; Kearns, Edwards, & Tingstrom, 1990) was also evaluated. The effect of using pre-selected \( p \) values to enter into the work sampling equations was also evaluated. Study 1 aimed to evaluate the utility of work sampling methods for behaviours of different durations by comparing extracted samples to a full time-of-interest dataset.

**METHOD**

**Participants and setting**

Jeremy, Toby, and Samuel were three males diagnosed with autism spectrum disorder and moderate to severe intellectual disabilities. Jeremy was a 6-year-old, who was non-verbal except for the word ‘no’. Toby was an 8-year-old with some imitative language and PECS manding, and Samuel was a 20-year-old with very little speech (imitated “yes” and “no” when prompted). All three attended a school for children with developmental and intellectual disabilities. Criteria for selection for the study included a requirement for regular attendance (so that a full week could be observed), and that the student had no history of absconding from the school grounds. All three participants engaged in at least two
inappropriate behaviours. With the exception of toilet stalls, observations were conducted across all areas of the school (classrooms, playground, swimming pool, etc.).

**Measurement and reliability**

Data were recorded as continuous duration measures by recording onsets and offsets using ObsWin32 (antam.co.uk) on laptop computers. Two adult behaviours and five participant behaviours were measured for each participant, and are operationally defined in Table 5.

**Table 5.**

*Operational Definitions for Each of the Adult and Participant Behaviours for Each of the Three Participants.*

<table>
<thead>
<tr>
<th>Category</th>
<th>Behaviour</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>Attention</td>
<td>Verbal comments (praise, reprimand, encouragement, statement, or question) directed to the student, gestures directed to the student (pointing, sign language, head shake), or physical contact initiated by the adult (e.g., prompts, pats, cuddles, handing something over).</td>
</tr>
<tr>
<td></td>
<td>Demand</td>
<td>Verbal instructions or requests directed to the student specifying that they were to commence, continue or cease a behaviour.</td>
</tr>
<tr>
<td>Participant</td>
<td>Positive affect</td>
<td>Vocalization or facial expression that could be deemed indicative of happiness (e.g., laughing, smiling, giggling) (Green &amp; Reid, 1996).</td>
</tr>
<tr>
<td></td>
<td>Negative affect</td>
<td>Vocalization or facial expression that could be deemed indicative of unhappiness (e.g., crying, screaming, wailing, yelling, frowning) (Green &amp; Reid, 1996).</td>
</tr>
<tr>
<td></td>
<td>On-task</td>
<td>Touching or manipulating materials related to the task in a manner as to complete the task, the student’s eyes oriented to the task materials or to the adult administering the task, following physical, gestural or verbal prompts related to task-completion, making comments or asking questions regarding the task, obtaining materials related to the task, and sitting or standing in the area in which task was to be completed.</td>
</tr>
<tr>
<td></td>
<td>Toe-walking (Jeremy)</td>
<td>Walking (placing one foot in front of the other) or standing on the toes or balls of his feet without his heels touching the floor.</td>
</tr>
</tbody>
</table>
EVALUATION OF WORK SAMPLING METHODS

<table>
<thead>
<tr>
<th>Behaviour</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grabbing (Jeremy)</td>
<td>Grasping one or both hands around another person’s clothing, hair or accessories (e.g., glasses, jewellery). Attempts (reaching one or both hands towards another person but was physically prevented from grabbing) included.</td>
</tr>
<tr>
<td>Grabbing (Toby)</td>
<td>Grasping one or both hands (or fingers) around another person’s body (arms, torso, head, neck). Sometimes accompanied by Toby pushing his head against, or hanging his body weight off the other person.</td>
</tr>
<tr>
<td>Tantrum (Toby)</td>
<td>Crying, slapping his head, stamping his feet, hitting surfaces, people or objects with an open hand, screaming or throwing objects. Unhappiness could progress to a tantrum when a second of the list of tantrum behaviours occurred for more than 5 s (in addition to vocalizations). Tantrums and indices of unhappiness were mutually exclusive.</td>
</tr>
<tr>
<td>Stereotypy (Samuel)</td>
<td>Arms out of sleeves and tucked inside sweater, sweater pulled up over head beyond chin level, finger picking (using the nail of one finger to scratch the skin of another), biting nails, and ritualistic or repetitive walking (e.g., one step forward, two steps back or walking in a shuffle.</td>
</tr>
<tr>
<td>Running (Samuel)</td>
<td>Distinguished from walking by the pace (gauged by the speed relative to the speed of other students and adults walking around him). Included running at and ducking around people and objects, and engaging in repetitive or ritualistic stepping whilst running (e.g., two step forwards, three steps back).</td>
</tr>
</tbody>
</table>

To calculate daily and weekly durations for each behaviour, the percentage of time spent in each behaviour was computed. Sessions were divided into 1-s intervals and the ObsWin32 variable statistics function reported the percentage of 1-s intervals in which each behaviour occurred. For MTS, percentage was calculated by dividing the number of MTS observations during which the behaviour was observed by the total number of MTS observations conducted in the session, and multiplying by 100. Hereafter, single momentary time samples (i.e., snap observations in work sampling) will be referred to as observations and extracted samples will be referred to as observation sessions.

Interobserver agreement (IOA) data were collected for 25.7% of the total observation time across all sessions and participants. Independent graduate student observers
experienced in data collection were provided with behavioural definitions prior to commencing observations. Practice observations, in which verbal agreement on examples of each behaviour was obtained, were conducted prior to independent observations. Exact agreement IOA was calculated for each behaviour by dividing recordings from an observation session into 10-s bins and computing the number of seconds of responding in each 10-s bin. The number of intervals in which exact agreement on seconds of occurrence or agreement on non-occurrence was obtained and divided by the total number of intervals. Mean IOA for attention was 83.8% (range, 68.4% to 96.2%), 94.2% for demand (range, 86.0% to 99.1%), 90.0% for on-task (range, 68.5% to 97.7%), 94.2% for positive affect (range, 72.5% to 100%), and 95.0% for negative affect (range, 88.9% to 98.5%). For Jeremy, mean IOA for the inappropriate behaviours was 91.9% for grabbing (range, 84.6% to 99.2%) and 82.8% for toe-walking (range, 82.1% to 83.5%). For Toby, mean IOA for the inappropriate behaviours was 99.7% for grabbing (range, 99.5% to 99.9%) and 99.2% for tantrums (range, 98.1% to 100%). For Samuel, mean IOA for the inappropriate behaviours was 98.3% for running (range, 97.5% to 99.1%) and 90.9% for stereotypy (range, 89.4% to 92.5%).

**Background activities.** The four background activities were *individual activity*, *group activity*, *free time*, and *transition*. The onset and offset of each background activity was denoted by a discriminative stimulus such as a bell ringing or a verbal statement from an adult that signalled onset or offset (e.g., a demand to sit down, presentation of task materials). Individual activities were defined as activities in which the students had their own task materials (e.g., blocks, pen and paper) and were allocated to separate areas (e.g., sitting separately at a table, or being in a particular area of the classroom such as the sink). Group activities were those in which the students were sharing task materials (e.g., a big book, parachute) and were seated or standing in a common area (e.g., circle time). Free time was
defined as a period in which students were permitted to play with leisure items and there was no structured activity in place. Transitions were the time between the cessation of an individual activity, group activity, or free time, and the onset of the next activity. Background activities were recorded as mutually exclusive categories in the ObsWin32, and were for the purposes of providing feedback to the school and parents. These activities were not included in the data analysis, except that on-task behaviour was not recorded during free time (free time periods were excluded from the analyses of on-task behaviour as, by definition, the participants could not be on-task or off-task during this time).

PROCEDURE

Preliminary observational recording. The purpose of preliminary observational recording was to obtain initial $p$ values to be entered into Equations 3 and 4. Preliminary observational recording was conducted in the two weeks prior to the full week of observational recording. The school day was divided into three blocks – 9.00 a.m. to 11.00 a.m., 11.00 a.m. to 1.00 p.m., and 1.00 p.m. to the end of the school day (between 2.30 p.m. and 3.00 p.m.). One observational recording, 60 min in duration, was conducted in each block for each participant, resulting in three preliminary observation sessions for each participant. Not all blocks were observed in one day; preliminary observational recording occurred across at least two days for each participant.

Teacher estimates of $p$. At the conclusion of each full week of observation for each participant, the researcher administered a verbal questionnaire to the classroom teacher. The researcher read the operational definition of each behaviour, and gave examples of inclusions and exclusions. Then the teacher gave an estimate of the percentage of time that the student spent engaged in each behaviour across a school week (Appendix 1). Teachers were informed that the percentages were not required to add up to 100% because students could be engaged
in several behaviours at once. In addition, the researcher asked teachers to identify a time of
day and time during the week that they would recommend as the best time for an observer to
record each of the behaviours.

**Week observational recording.** Full week observational recording was conducted
for each participant (five observation days for Jeremy and Toby and four observation days for
Samuel). No data were recorded on the Wednesday of the observation week for Samuel
because he went off-site for work experience every Wednesday, and this was not considered
to be part of the time-of-interest. No data recording was conducted between 11.30 a.m. and
12.30 p.m. on the Thursday of the observation week for Samuel because the class left the
school grounds for swimming in a community pool. The onset of an observation occurred
with the ringing of the school bell (at 9.00 a.m.), and recording was terminated at the end of
the day when each student exited the classroom. Observations days were between 14,495 s
(approximately four hours) and 20,659 s (approximately five and three quarter hours) in
duration.

**Work sampling equations.** Although in the equation reported by Barnes (1957), $p$ is
a proportion of time that the behaviour occurs, from this point forth, $P_O$ (percentage of time
spent engaged in the behaviour) is substituted for $p$ to facilitate comparison with reported
measures of duration in behaviour analysis.

\[
N = \frac{4 (100 - P_O)}{S^2 P_O} \quad (3)
\]

\[
N = \frac{4 P_O (1 - P_O)}{S^2} \quad (4)
\]

The mean percentage of time spent in each behaviour across the three preliminary
observation sessions was entered as the initial $P_O$ value in Equations 3 and 4 (using relative
and absolute accuracy, respectively). For both Toby and Samuel, one behaviour was not
observed during preliminary observation sessions (*tantrum* and *negative affect*, respectively). For these two behaviours, a $P_O$ value of 1% was used in the initial calculations. The desired accuracy was ± 5% ($S$) for both equations. Using these values, the number of required observations was calculated for each behaviour for each participant. The number of required observations was then divided by the number of days over which observations were conducted (five for Jeremy and Toby, four for Samuel) in order to determine the number of observations to be conducted each day. After each observation session, the obtained $P_O$ was entered into the equation. The number of observations already extracted was subtracted from the new $N$ (required number of observations), and the remainder divided by the number of remaining observation days. If a behaviour had not been observed in the observation session (the new $P_O$ value was 0), the mean $P_O$ value of all observation sessions was used.

The following analyses were conducted for each behaviour across all participants (Figure 9). Equation 3 (relative accuracy) with preliminary observation-based estimates of $P_O$ (referred to henceforth as Prelim-R); Equation 3 with staff estimates of $P_O$ (Staff-R); Equation 4 (absolute accuracy) with preliminary observation-based estimates and staff estimates of $P_O$ (Prelim-A, and Staff-A, respectively). In addition, each of these four analyses was conducted using each of the sampling methods (three discontinuous sampling methods and one continuous sampling method commencing at four different times of day; seven sampling methods in total). As a result, 224 combinations of variables (four equation combinations by eight categories by seven sampling methods) were evaluated for Jeremy and Toby, and 196 combinations for Samuel (fewer combinations due to one fewer observation day and one behaviour which was never observed and was eliminated from the analyses).
Figure 9. Flowchart showing each of the analyses conducted (total of 28 combinations shown). Each of the combinations of analyses was conducted for each behaviour, resulting in 224 analyses for two participants (eight behaviours), and 196 analyses for the third participant (seven behaviours).
**Momentary time sampling (random).** A random sequence generator available from www.random.org was used to place each number from 1 to the end of the day (in seconds) in a random order. The timing of each observation was selected by using the numbers generated in the sequence. For example, if 300 observations were required, the first 300 numbers in the sequence were used. The sequence generator sampled without replacement. Each behaviour was identified as occurring or not occurring at each MTS observation time by comparing the observation times to a printed list of all the seconds in which the behaviour occurred. The percentage of time spent engaged in each behaviour was calculated by dividing the number of observations in which the behaviour was observed by the total number of observations.

**Momentary time sampling (fixed-interval).** The total duration of the day in seconds (e.g., 20,607 s) was divided by the number of observations required per day by Equation 3 or 4 (e.g., 80), and the resulting quotient rounded down to the nearest whole number (e.g., 257 s). The resulting number represented the time (in seconds) between observations. Using the sampling interval conversion function in ObsWin32, the data file for each day was converted into a momentary time sample file with interval duration calculated as described above. The converted file then showed the number of MTS observations during which the behaviour of interest occurred.

**Momentary time sampling (non-continuous).** Each day was divided into intervals of equal duration by the method described for fixed-interval sampling. MTS observations were conducted once per interval, but were randomly allocated within each interval. A number between 0 and the value of the interval duration was selected randomly for each interval using the random integer generator from www.random.org. Each random number was added to each interval onset time (in seconds), and this represented the second at which each MTS observations was to be conducted. As in random sampling, each behaviour was identified as occurring or not occurring at each observation time by comparing the
observation times to a printed list of all the seconds in which the behaviour occurred. The percentage of time spent in each behaviour was calculated by the number of MTS observations in which the behaviour occurred by the total number of MTS observations.

**Continuous sampling.** The number of observations required by the equations was conducted across continuous seconds. For example, if an equation calculated that 1,200 observations were to be conducted on the first day, 1,200 consecutive seconds (20 min) were extracted. Samples were extracted starting from 9.00 a.m., 11 a.m., 1 p.m., or a randomly-selected time. Random times were selected by using www.random.org to select a random number between 0 and the end of the day in seconds (e.g., 20,590). Extracted observation sessions were required to be a minimum of 600 s (10 min), because this was determined to be the minimum duration of observations likely to be conducted in natural settings. Observations were extracted by using the time filter function in ObsWin32 to specify the duration and onset of each observation. The percentage of time spent in each behaviour was then obtained from the summary statistics.

For on-task behaviour, if the entirety of the random time selected for observation fell within a period of *free time*, it was discarded and another selected. If the duration of the required observation exceeded the remainder of the session, the observation duration was recorded as the time between the onset of the observation and the end of the day. This resulted in some extracted observations being shorter than specified by the equation. If the calculated number of observations required on the next day was less than the total number of observations already conducted (and therefore was negative), the same number of observations as the previous day was conducted.

True values of 0 (i.e., behaviours that did not occur at all on that day) were removed from the analysis (*negative affect* was never observed for Samuel and *tantrum* was not observed on four days out of five for Toby). Removing these data distinguished instances in
which behaviours did occur but were not sampled, from days in which behaviours did not occur and thus could not be sampled.

**Summary reports.** The school, teacher, and parents of each participant were provided with a summary of the percentages of time the student spent engaging in each of the behaviours. Each summary report also included some suggestions for increasing on-task behaviour (identified as a deficit for all three participants), and general suggestions for addressing the challenging behaviour. All suggestions were supported by the data collected and qualitative notes recorded by the researcher during observations (using the notes function in ObsWin32).

**ADDITIONAL ANALYSES**

**Observation time.** The researcher selected one behaviour for each participant based on teacher responses to the questionnaire (Table 6). For each behaviour, the teacher specified a particular time of day during which they predicted the behaviour could be observed (Appendix 1). A 3,600-s (60-min) continuous observation session was extracted at the nominated time of day for each day of the week and compared to daily true values. For Participant 2, the teacher specified an activity rather than a time of day and so samples were extracted starting at the commencement of this activity, and finished at the termination of the activity. As a result, the samples extracted in this analysis for Participant 2 were less than 3,600 s in duration (range 1,883 s to 2,659 s).

**Rule of 1,000.** The 1,000 required observations were divided by the number of days over which the study was conducted (five days for Jeremy and Toby, four for Samuel, resulting in 200 and 250 per day, respectively). The total duration of each day (in seconds) was divided by 200 or 250, and the resulting quotient rounded down to the nearest whole
number. Samples were extracted using the same method for extracting momentary time samples as described above.

**Five-minute momentary time sampling.** Momentary time samples were extracted every five minutes (e.g., 300 s) using the sampling interval conversion function in ObsWin32 (as described above). The total number of observations conducted across the week was 319, 327, and 247 for Participants 1, 2, and 3 respectively.

**Varying $P_O$ values.** The researcher selected randomly at least one behaviour from each participant for the varying $P_O$ value analysis (Table 6). For each of the selected behavioural categories, $P_O = 10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, \text{and } 90\%$ were used as the initial $P_O$ estimate in Equation 3. Samples were extracted using the fixed-interval discontinuous sampling method. Obtained $P_O$ values were not entered into the equations to adjust the number of required MTS observations: the number conducted was specified by the initial $P_O$ estimate and divided equally across the number of days of the study.

Table 6.
*Behaviours Selected and Used in Additional Analyses in Study 1.*

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Duration</th>
<th>Observation time</th>
<th>$p$ value analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt; 0%, &lt; 1%</td>
<td>&quot;</td>
<td>Participant 3 Demand</td>
</tr>
<tr>
<td></td>
<td>1% – 4.9%</td>
<td>Participant 3 Running</td>
<td>Participant 2 Negative affect</td>
</tr>
<tr>
<td></td>
<td>5% – 9.9%</td>
<td>&quot;</td>
<td>Participant 1 Negative affect</td>
</tr>
<tr>
<td></td>
<td>10% – 29.9%</td>
<td>Participant 1 On-task</td>
<td>Participant 3 On-task</td>
</tr>
<tr>
<td></td>
<td>30% – 43.5%</td>
<td>Participant 2 On-task</td>
<td>Participant 1 Attention</td>
</tr>
</tbody>
</table>
RESULTS

Individual participants’ behaviours were grouped into five windows by the overall percentage of time in which they were observed across the week, so that comparisons could be made between high-duration (30% - 43.5%), intermediate-duration behaviours (1% - 4.9%, 5% – 9.9%, and 10% - 29.9%), and low-duration behaviours (< 1%). The highest-duration behaviour for any of the participants occurred for 43.5% of the week (stereotypy for Samuel). Each window contained between three and six behaviours from at least two of the three participants (Table 7). True values below the minimum window size resulted from true daily percentages that were smaller than the overall weekly percentage ($P_w$) from which the window values were selected.

Table 7.
Behaviours Measured for Each of the Three Participants Grouped into Windows by Overall True Weekly Duration (Overall Percentage of the Week Spent Engaged in the Behaviour).

<table>
<thead>
<tr>
<th></th>
<th>&gt; 0%, &lt; 1%</th>
<th>1% – 4.9%</th>
<th>5% – 9.9%</th>
<th>10% – 29.9%</th>
<th>30% – 43.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jeremy</td>
<td>Demand</td>
<td>Negative</td>
<td>On-task</td>
<td>Attention</td>
<td>Toe-walking</td>
</tr>
<tr>
<td></td>
<td></td>
<td>affect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grabbing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>affect</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Toby</td>
<td>Grabbing</td>
<td>Demand</td>
<td>Attention</td>
<td>On-task</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tantrum</td>
<td>Negative</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>affect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>affect</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Samuel</td>
<td>Demand</td>
<td>Running</td>
<td>Attention</td>
<td>On-task</td>
<td>Stereotypy</td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>affect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DATA PRESENTATION ON REPRESENTATIVENESS

For each extracted observation session, the percentage of time in which a behaviour was observed (obtained percentage of time, $P_O$) was plotted against the overall percentage of time in which that behaviour occurred for the day ($P_D$). Figures 10 and 11 compare the true percentages of behaviours with the percentages obtained with numbers of observations selected with Equations 3 and 4. Numbers of observations extracted using MTS are presented on the left panels, and using continuous sampling on the right panels.

Figures 10 and 11 show a line representing perfect correspondence between obtained and true values. The closer the data points are to the perfect correspondence line, the more representative they are. Figures 10 and 11 also show the relative accuracy limits as specified by Equation 3 (converging dashed lines plotted by calculating $P_O \pm 5\%$ of $P_O$ for values of $P_O$ from 10\% to 100\%) and Equation 4 (parallel dashed lines plotted by calculating $P_O \pm 5\%$ for values of $P_O$ from 10\% to 100\%). When the number of observations required by each equation for behaviours of different durations (i.e., different $P_O$ values) is calculated for a desired accuracy of $\pm 5\%$, 95\% of data should fall within the accuracy limits (i.e., will be representative). It is important to note that the lines plotted in Figures 8 and 9 are not regression lines with confidence intervals (i.e., are not calculated from the data).

Figures 10 to 13 include all momentary time sampling (fixed, random, and non-continuous) and all continuous sampling (commencing at 9 a.m., 11 a.m., 1 p.m., and at mixed times) data. The effects of these variables are addressed in subsequent figures.

Momentary time sampling using Equation 3 (relative accuracy). When momentary time sampling was used to extract the number of observations required by Equation 3, 80.9\% of the obtained percentages fell within the $\pm 5\%$ relative accuracy limits of the true percentage (Figure 10), demonstrating high representativeness. Low-duration behaviours were sampled more representatively than higher-duration behaviours (i.e., 77.2\%
and 90% of the obtained percentages fell within the ± 5% relative accuracy limits for
behaviours occurring for ≥ 30% and < 1% of the week, respectively). Behaviour was
detected in all 99.6% of observation sessions (i.e., there were two observation sessions in
which $P_O = 0$). Table 8 shows the percentage of data points that fell within the accuracy
limits and the percentage of observations sessions in which $P_O = 0$ for Figures 10 and 11.

Table 8.  
Percentage of Observation Sessions in which Obtained Percentages ($P_O$) fell within the ± 5% Accuracy Limits of the True Percentage ($P_D$), and the Percentage of Observation Sessions in Which Behaviour was not Detected ($P_O = 0$). Behaviours were Grouped into Windows by Overall True Weekly Duration and the Total Percentages (Across All Behaviours) are Presented in the Far Right Column. Equation 3 used Relative Accuracy Limits and Equation 4 used Less-Stringent Absolute Accuracy Limits.

<table>
<thead>
<tr>
<th></th>
<th>&gt;0%, &lt; 1%</th>
<th>1% – 4.9%</th>
<th>5% – 9.9%</th>
<th>10% – 29.9%</th>
<th>30% – 43.5%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTS Equation 3 limits</td>
<td>90%</td>
<td>81.0%</td>
<td>80.8%</td>
<td>76.2%</td>
<td>77.2%</td>
<td>80.9%</td>
</tr>
<tr>
<td>% of 0s</td>
<td>0%</td>
<td>0%</td>
<td>1.28%</td>
<td>1.19%</td>
<td>0%</td>
<td>.37%</td>
</tr>
<tr>
<td>MTS Equation 4 limits</td>
<td>100%</td>
<td>87.4%</td>
<td>75.6%</td>
<td>65.5%</td>
<td>83.3%</td>
<td>83.5%</td>
</tr>
<tr>
<td>% of 0s</td>
<td>95.6%</td>
<td>56.3%</td>
<td>14.1%</td>
<td>3.6%</td>
<td>0%</td>
<td>36.7%</td>
</tr>
<tr>
<td>Continuous Equation 3 limits</td>
<td>19.2%</td>
<td>12.9%</td>
<td>7.7%</td>
<td>4.5%</td>
<td>8.6%</td>
<td>11%</td>
</tr>
<tr>
<td>% of 0s</td>
<td>25.8%</td>
<td>7.8%</td>
<td>9.6%</td>
<td>20.5%</td>
<td>13.8%</td>
<td>14.3%</td>
</tr>
<tr>
<td>Continuous Equation 4 limits</td>
<td>97.5%</td>
<td>80.2%</td>
<td>40.4%</td>
<td>7.1%</td>
<td>4.6%</td>
<td>50%</td>
</tr>
<tr>
<td>% of 0s</td>
<td>84.2%</td>
<td>50.9%</td>
<td>55.8%</td>
<td>42.9%</td>
<td>23.0%</td>
<td>50%</td>
</tr>
</tbody>
</table>
Figure 10. Representativeness of extracted momentary time samples (fixed, random, and discontinuous) shown as the obtained percentage of time spent in each behaviour ($P_O$) plotted against the overall daily percentage of time spent in that behaviour ($P_D$). Graphs show representativeness for different windows of true durations from high (30% - 43.5%) to lowest (< 1%).
Momentary time sampling using Equation 4 (absolute accuracy). When momentary time sampling was used to extract the number of observations required by Equation 4, 83.9% of the obtained percentages fell within the less stringent ± 5% absolute accuracy limits of the true percentage (Figure 10). As when Equation 3 was used, more of the obtained percentages for low-duration behaviours fell within ± 5% absolute accuracy limits than high-duration behaviours (e.g., 83.3% and 100% of the obtained percentages for behaviours occurring for ≥ 30% and < 1% of the day, respectively). When Equation 4 was used, behaviour was more likely to be undetected than when Equation 3 was used (i.e., PO = 0 was obtained in 36.7% of observation sessions). Behaviour was also less likely to be detected in low-duration behaviours. For example, PO = 0 was obtained in 0% of observation sessions of high-duration behaviours (occurring for ≥ 30% of the day) and 95.6% of observation sessions of low-duration behaviours (occurring for < 1% of the day) (Table 8).

Continuous sampling using Equation 3 (relative accuracy). When continuous sampling was used to extract the number of observations required by Equation 3, obtained percentages of higher-duration behaviours were less representative than when MTS was used (Figure 11). True percentages were both underestimated and overestimated (e.g., top left panel, Figure 11). With the use of continuous sampling, 11% of the obtained percentages from extracted observation sessions fell within the ± 5% relative accuracy limits, demonstrating low representativeness. As the overall duration of the behaviour decreased, data were closer to the line showing perfect correspondence (i.e., were more representative), but most still fell outside the ± 5% accuracy limits (only 19.2% of the data points fell within the relative accuracy limits for behaviours occurring < 1% of the week). The behaviour of interest was not observed in the observation session (PO = 0) in a mean of 14.3% of the data points (range of 7.8% and 25.8% across windows).
Figure 11. Representativeness of extracted continuous samples shown as the obtained percentage of time spent in each behaviour ($P_O$) plotted against the overall daily percentage of time spent in that behaviour ($P_D$). Graphs show representativeness for different windows of true durations from high (30% - 43.5%) to lowest (< 1%).
Continuous sampling using Equation 4 (absolute accuracy). When continuous sampling was used to extract the number of observations required by Equation 4, obtained percentages of higher-duration behaviours similarly poorly representative to when Equation 3 was used (Figure 11). Percentages were both under- and overestimated, but for behaviours occurring for 1% to 4.9% and < 1% of the week, the majority of data points underestimated the duration of the behaviour. Of the obtained percentages from extracted observation sessions, 50% fell within the ± 5% absolute accuracy limits. However, behaviour was not detected ($P_O = 0$) in 50% of all extracted observation sessions, with many sessions of $P_O = 0$ obtained for low-duration behaviours (e.g., 84.2% of extracted sessions for behaviours occurring < 1% of the week). Therefore, although 50% of the obtained percentages were inside the ± 5% absolute accuracy limits, many of the data points that were $P_O = 0$ were included because the limits included 0.

DATA PRESENTATION ON EFFORT

Figures 12 and 13 show the number of MTS observations and continuous sampling seconds of observation (left and right panels, respectively) plotted against the absolute difference between the obtained and true percentages (calculated by subtracting $P_D$ from $P_O$). These data show the effort (i.e., the number of observations) required to obtain representative samples. Data points that lie closer to 0 on the x-axis (dashed line) are more representative of true values than those further away.

Momentary time sampling using Equation 3 (relative accuracy). Figure 12 shows that the almost perfect correspondence for low-duration behaviours (Figure 10) was due to the larger number of observations conducted for behaviours of lower duration, with some observation sessions containing over 20,000 MTS observations (one per second, corresponding to continuous observation). Fewer MTS observations were required by
Equation 3 for higher-duration behaviours (e.g., ≥ 30% and 10% - 29.9%), but small absolute differences between obtained and true percentages were observed (Figure 12). Smaller numbers of observations also produced representative data for behaviours of intermediate durations (e.g., middle left panel, Figure 12).
Figure 12. Effort (number of observations in each extracted observation session) plotted against the difference between the obtained percentage of time spent in a behaviour ($P_O$) and the overall daily percentage of time spent in that behaviour ($P_D$) for momentary time samples. Graphs show representativeness for different windows of true durations from high (30% - 43.5%) to lowest (< 1%).
Momentary time sampling using Equation 4 (absolute accuracy). Figure 12 also shows that Equation 4 required fewer observations than Equation 3 (the maximum number of observations in any session was 449). Larger absolute differences between obtained and true percentages were obtained when MTS and Equation 3 were used, and larger absolute differences were obtained for higher-duration behaviours (e.g., occurring for ≥ 30% of the day).

Continuous sampling using Equation 3 (relative accuracy). Figure 13 shows that large numbers of observations were required when Equation 3 was used to select continuous observation session durations, particularly for low-duration behaviours (i.e., occurring for < 1% of the day, bottom left panel). Smaller numbers of observations resulted in larger absolute differences between obtained and true percentages for higher duration behaviours (e.g., top left panel), but not for low-duration behaviours (e.g., occurring for < 1% of the day, bottom left panel). Some larger numbers of observations also resulted in large absolute differences (and therefore, poorly representative samples) (e.g., top left panel).

Continuous sampling using Equation 4 (absolute accuracy). As each continuous observation session was required to be a minimum of 600 s (10 min) in duration, all observation sessions selected using Equation 4 were 600 s (because the number required by the equation was always less than 600 s). Figure 13 shows that for higher-duration behaviours (e.g., occurring for ≥ 10% of the day), continuous observation sessions 600 s in duration resulted in a wide range of absolute differences between obtained and true percentages (e.g., top right panel). More representative samples (i.e., smaller differences between obtained and true percentages) were obtained for lower-duration behaviours (e.g., bottom right panel).
Figure 13. Effort (number of observations in each extracted observation session) plotted against the difference between the obtained percentage of time spent in a behaviour ($P_O$) and the overall daily percentage of time spent in that behaviour ($P_D$) for continuous samples. Graphs show representativeness for different windows of true durations from high (30% - 43.5%) to lowest (< 1%).
Use of work sampling equations. When Equation 3 was used for both MTS and continuous sampling, some observation sessions contained numbers of observations that were outside the number of predicted observations shown on Figure 3 (Chapter 2). For example, Figure 3 shows that 14,400 observations are required when $P_O$ is 10%, and therefore that 2,880 observations would be required on each of five days of the study. However, the number of observations conducted in each session for behaviours occurring between 5% and 10% of the day ranged from 80 to 20,659 (Figures 12 and 13). Figure 4 shows that 144 observations are required by Equation 4 when $P_O$ is 10% (29 observations per day across five days). However, when Equation 4 was used to extract MTS samples for behaviours occurring between 5% and 10% of the day, a range of 2 to 200 observations were conducted (Figure 12, right panel). Due to the changing value of $P_O$ that was entered daily into Equations 3 and 4, the number of required observations varied across days. Changes to the required number of observations ($N$) were large if differences between $P_O$ values obtained across subsequent days were large. Also, inaccurate preliminary estimations of $P_O$ resulted in required $N$ that differed greatly from the number of observations predicted by Figures 3 and 4.

As the observation samples extracted using Equation 4 (absolute accuracy) produced poorly representative data, data extracted using Equation 4 were removed from subsequent analyses. From this point forth, only data samples extracted using Equation 3 (relative accuracy) are included in the figures.

EFFECT OF SAMPLING SCHEDULE

Figure 14 shows the effort (number of observations required) for the three methods of scheduling momentary time samples, compared to the effort required to conduct continuous samples. There was no difference between the effort required to conduct the different schedules of MTS, and no difference in the range of absolute differences between obtained
and true percentages (i.e., similarly representative samples were obtained). Continuous samples produced a broader range of absolute differences between obtained and true percentages (-64.3% to 66.2%), with large absolute values indicating poor representativeness. Longer continuous samples produced more representative samples (i.e., data points were close to 0 on the x-axis), as indicated by the solid horizontal lines (Figure 14, bottom right panel). More representative samples were obtained when continuous observation sessions were three hours or longer in duration.

Figure 14. Effort (number of observations in each extracted observation session) plotted against the difference between the obtained percentage of time spent in a behaviour \( (P_O) \) and the overall daily percentage of time spent in that behaviour \( (P_D) \) for the three schedules of momentary time sampling and continuous sampling. The solid horizontal lines indicate the duration of continuous samples in hours to the equivalent number of observations.
Figure 15 shows that regardless of whether continuous samples commenced at 9 a.m., 11 a.m., 1 p.m., or at a randomly-selected time (mixed), behaviours were not sampled representatively in the majority of observation sessions. Of the obtained percentages, 11.7% fell within relative accuracy limits (range of 2.2% to 27.8% across the four schedule timings). As continuous observation samples were shown to produce poorly representative data, from this point forth, only MTS samples are included in the figures and analyses unless otherwise indicated.

*Figure 15.* Representativeness of extracted continuous samples commencing at differently scheduled times shown as the obtained percentage of time spent in each behaviour ($P_O$) plotted against the overall daily percentage of time spent in that behaviour ($P_D$).
VALUES PREDICTED BY STAFF AND PRELIMINARY OBSERVATION ESTIMATES

Figure 16 shows that there was little difference between the effort (i.e., the number of required observations) and the resulting absolute differences between obtained and true percentages for numbers of observations required by Equation 3 when a staff estimate or preliminary percentage was used as the initial $P_O$ value. Due to the changing value of $P_O$ that was entered daily into Equation 3, inaccurate initial $P_O$ values affected only the first observation session for each behaviour.

![Figure 16](image)

*Figure 16.* Effort (number of observations in each extracted observation session) plotted against the difference between the obtained percentage of time spent in a behaviour ($P_O$) and the overall daily percentage of time spent in that behaviour ($P_D$) for continuous samples. The data are from samples extracted using Equation 3 and fixed momentary time sampling.

Preliminary observation sessions yielded more accurate estimates of $P_O$ than did values predicted by staff (Figure 17). Staff consistently overestimated the percentage of time participants spent engaging in all but two behaviours, and were more likely to overestimate the percentage of time spent engaging in lower-duration behaviours than higher-duration behaviours. Absolute differences between estimated ($P_E$) and true weekly percentages ($P_W$)
were large in some instances (e.g., a difference of 81% for Participant 3, bottom left panel). Preliminary observations produced both under- and overestimates of $P_0$, but produced more accurate estimates of $P_W$ for lower-duration behaviours (occurring for $\leq 20\%$ of the day). Absolute differences between estimated and true percentages ($P_E - P_W$) were similar across all participants (i.e., all staff members overestimated $P_W$ for most behaviours).

**Figure 17.** The difference between the weekly percentage of time spent in a behaviour ($P_W$) and the estimated percentage of time spent in that behaviour ($P_E$) as determined by teacher estimations (closed circles) or preliminary observations (open circles).

**OBSERVATION TIME ANALYSIS**

Figure 18 shows that when continuous observations were extracted at teacher-nominated times, 33.3% of the samples were representative (i.e., fell within $\pm 5\%$ relative
accuracy limits). Each of the obtained percentages within ± 5% relative accuracy limits were for higher-duration behaviours (e.g., behaviours occurring for ≥ 10% of the day). Running (Participant 3) was always overestimated (Figure 18, bottom left panel), but almost perfect correspondence between obtained and true percentages was obtained in two samples (top panels). There was one sample in which behaviour was not detected (i.e., \( P_O = 0 \) ) (top right panel).

Figure 18. Representativeness of continuous samples extracted at times nominated by the teachers shown as the obtained percentage of time spent in each behaviour (\( P_O \)) plotted against the overall daily percentage of time spent in that behaviour (\( P_D \)).
Figures 19, 20, and 21 show the distribution of on-task behaviour (Participants 1 and 2) and running behaviour (Participant 3) across the days of the week. The dashed-edge boxes show the timing and duration of the extracted continuous observation samples for which the data are presented in Figure 18. Figures 19, 20, and 21 show that teachers were able to identify times of the day during which the behaviour was likely to occur.

**Figure 19.** Occurrence graph of on-task behaviour for Participant 1. Black bars show the occurrence (location) and duration (width) of each instance of running across each of the five days of the observed week. The dashed-edge box indicates the observation time nominated by the teacher.

**Figure 20.** Occurrence graph of on-task behaviour for Participant 2. Black bars show the occurrence (location) and duration (width) of each instance of running across each of the five days of the observed week. The dashed-edge boxes indicate the observation times nominated by the teacher.
EVALUATION OF WORK SAMPLING METHODS

Figure 21. Occurrence graph of running behaviour for Participant 3. Black bars show the occurrence (location) and duration (width) of each instance of running across each of the four days of the observed week. The dashed-edge box indicates the observation time nominated by the teacher.

VARYING \( P_O \) VALUES

When \( P_O = 10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, \) and 90\% were used as the initial \( P_O \) estimate in Equation 3 to select the number of momentary time samples, more representative samples were obtained when more observations were conducted (smaller \( P_O \) values) and for behaviours of higher duration (Figure 22). More obtained percentages fell within ± 5\% relative accuracy limits for higher-duration behaviours (e.g., 71.1\% of data points for behaviours occurring for ≥ 30\% of the week) than for lower-duration behaviours (e.g., 11.1\% of data points for behaviours occurring for <1\% of the week) (Table 9).

Table 9.
Total Percentage of Observation Sessions in which Obtained Percentages (\( P_O \)) fell within the ± 5\% Accuracy Limits of the True Percentage (\( P_D \)) across Behaviours of Different Weekly Duration (Windows) and when Varying \( P_O \) Values were used.

<table>
<thead>
<tr>
<th>Bin</th>
<th>&gt; 0%, &lt; 1%</th>
<th>1% – 4.9%</th>
<th>5% – 9.9%</th>
<th>10% – 29.9%</th>
<th>30% – 43.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>11.1%</td>
<td>37.8%</td>
<td>42.2%</td>
<td>91.7%</td>
<td>71.1%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( P_O ) value</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>87%</td>
<td>70%</td>
<td>56.5%</td>
<td>65.2%</td>
<td>56.5%</td>
<td>39.1%</td>
<td>30.4%</td>
<td>34.8%</td>
<td>17.4%</td>
</tr>
</tbody>
</table>
Larger number of observations conducted for smaller $P_O$ values also resulted in a higher percentage of obtained percentages falling within ± 5% relative accuracy limits (e.g., 87% and 17.4% of data points for behaviours occurring for $\geq 30\%$ and $< 1\%$ of the week, respectively).

*Figure 22.* Representativeness of extracted fixed momentary time samples shown as the obtained percentage of time spent in each behaviour ($P_O$) plotted against the overall daily percentage of time spent in that behaviour ($P_D$). The duration of each sample was determined by entering a $P_O$ value of 10% to 90% into Equation 3. Darker data points represent smaller $P_O$ values and therefore larger numbers of observations.
CALIBRATION ANALYSIS

For the samples obtained using the Rule of 1,000 and 5-min MTS, true daily percentages and the total number of observations conducted across the week ($P_D$ and $N$ respectively) were entered into Equation 3 to calculate the likely accuracy, $S$. For example, the true daily percentage of on-task behaviour for Participant 1 on Thursday was 13.4% and 317 MTS were conducted at 5-min intervals across the week, resulting in predicted $S = \pm 28.5\%$. The second column from the right in Table 10 shows the mean accuracy (S) for each behaviour bin. Low-duration behaviours produced samples with much poorer accuracy than higher durations samples (e.g., 95% of the data points obtained under the Rule of 1,000 were likely to fall within $\pm 123.1\%$ and $\pm 8.5\%$ of the true daily percentages for behaviours occurring for $< 1\%$ and $\geq 30\%$ of the week respectively). Calculated accuracy values were higher for the Rule of 1,000 because more observations were conducted than when 5-min MTS was used.

To calculate the upper and lower relative accuracy limits, $S$ was multiplied by $P_D$ and the resulting quotient subtracted from (lower limit) and added to (upper limit) $P_D$. To continue with the above example of on-task behaviour on Thursday for Participant 1, 28.5% was multiplied by 13.4%, and then subtracted from and added to 13.4%. The resulting limits predicted that 95% of the data should fall within 9.6% and 17.2%. The far right in Table 10 shows the percentage of data points in each behaviour window that fell within the calculated accuracy limits. Overall, a higher percentage of data points fell within the calculated accuracy limits for lower-duration behaviours. A higher percentage of data points also fell within the calculated accuracy limits when fewer observations were conducted and therefore accuracy limits were broader (5-min MTS).
Table 10. Linear Regression Statistics and Equation 3 Error Analysis for the Rule of 1,000 and 5-min Momentary Time Sample Data.

<table>
<thead>
<tr>
<th>Behaviour window</th>
<th>Linear regression statistics</th>
<th>Equation 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Slope</td>
<td>$R^2$</td>
</tr>
<tr>
<td>Rule of 1,000</td>
<td>&gt; 0%, &lt; 1%</td>
<td>0.542</td>
</tr>
<tr>
<td></td>
<td>1% – 4.9%</td>
<td>1.051</td>
</tr>
<tr>
<td></td>
<td>5% – 9.9%</td>
<td>1.118</td>
</tr>
<tr>
<td></td>
<td>10% – 29.9%</td>
<td>0.929</td>
</tr>
<tr>
<td></td>
<td>30% – 43.5%</td>
<td>0.946</td>
</tr>
<tr>
<td>5-min MTS</td>
<td>&gt; 0%, &lt; 1%</td>
<td>1.628</td>
</tr>
<tr>
<td></td>
<td>1% – 4.9%</td>
<td>3.284</td>
</tr>
<tr>
<td></td>
<td>5% – 9.9%</td>
<td>0.898</td>
</tr>
<tr>
<td></td>
<td>10% – 29.9%</td>
<td>0.801</td>
</tr>
<tr>
<td></td>
<td>30% – 43.5%</td>
<td>0.861</td>
</tr>
</tbody>
</table>

Calibration by linear regression has been shown to be useful to assess the quality of direct observation measures (Mudford, Zeleny, Fisher, Klum, & Owen, 2011). The researcher used calibration methods to evaluate the quality of the Rule of 1,000 and 5-min MTS measures of the true percentage of time ($P_D$) for each behaviour. A slope of 1.0 indicates that the obtained percentage of time ($P_O$) is an accurate measure of the criterion ($P_D$). A slope of less than 1.0 indicates a tendency to underestimate $P_D$ and a slope of greater than 1.0 indicates a tendency to overestimate $P_D$. $R^2$, a measure of the distance of each data point from the regression line, shows how much variability in the data can be explained. An $R^2$ value of 1.0 indicates that all the variability in the data can be explained by the linear model. A standard error of the estimate ($SE$) close to zero indicates precise measurement of
$P_D$ and less random error. The precision of the obtained values (expressed as ± percentage of the day) can be calculated from dividing the width of the precision intervals by 2. Table 10 shows the regression statistics for each behaviour window for the Rule of 1,000 and 5-min MTS analyses. The least squares method of linear regression was conducted in SigmaPlot 11.0 (Systat Software Inc.), and the regression line (solid) and prediction intervals (dashed) plotted in Figures 23 and 24.

**Rule of 1,000.** Figure 23 compares the obtained and true percentages of the behaviours for samples obtained through the Rule of 1,000. Low-duration behaviours (occurring for < 1% of the week) were likely to be underestimated (slope = 0.542), whereas higher-duration behaviours were more likely to be measured accurately (i.e., slopes closer to 1.0). Lower-duration behaviours were measured more precisely than higher-duration behaviours (i.e., produced small $SE$ values and narrower prediction intervals).

When Rule of 1,000 was used, 34.8% of the obtained percentages fell within the ± 5% relative accuracy limits (Figure 23). However, many of the other data points for behaviours occurring for ≥ 5% of the week were close to the ± 5% relative accuracy limits (top and middle panels). None of the obtained percentages for behaviours occurring for ≤ 4.9% of the week fell within the ± 5% relative accuracy limits. Of the samples extracted for behaviours occurring for < 1% of the week, 100% did not detect behaviour (i.e., $P_D = 0$). Therefore, more representative samples were obtained for higher-duration behaviours than lower-duration behaviours when the Rule of 1,000 was used to extract momentary time samples.
Figure 23. Representativeness of fixed momentary time samples extracted using the Rule of 1,000 shown as the obtained percentage of time spent in each behaviour ($P_O$) plotted against the overall daily percentage of time spent in that behaviour ($P_D$). Each data point represents either 200 (Participants 1 and 2) or 250 (Participant 3) observations. The regression lines are solid and the prediction intervals are the dashed lines.

**Five-minute momentary time sampling.** Figure 24 compares the obtained and true percentages of the behaviours for samples obtained through conducting momentary time samples every 5 min across the whole day. The number of momentary time samples per day...
ranged from 48 to 68 (affected by the duration of the day). Low-duration behaviours (occurring for < 1% and 1% - 4.9% of the week) were likely to be overestimated (slope = 1.628 and 3.284 respectively), whereas higher-duration behaviours (occurring for ≥ 5% of the week) were likely to be underestimated, but to a smaller degree (e.g., slope = 0.801 for the 10% - 29.9% bin). The lowest-duration behaviours were most precisely measured (SE = .009).

Of the obtained percentages, 8.9% fell within the ± 5% relative accuracy limits. Obtained percentages of higher-duration behaviours (i.e., occurring for ≥ 30% of the week) were more likely to fall within the ± 5% relative accuracy limits than lower-duration behaviours. For example, 21.1% and 3.4% of the obtained percentages fell within the ± 5% relative accuracy limits for behaviours occurring for ≥ 30%, and 5% to 9.9% of the week (respectively). None of the obtained percentages for behaviours occurring for < 1% of the week fell within the ± 5% relative accuracy limits. Behaviour was not detected ($P_o = 0$) in 24.4% of observation sessions, most of which were sessions for the lowest-duration behaviours.
Figure 24. Representativeness of 5-min fixed momentary time samples shown as the obtained percentage of time spent in each behaviour ($P_O$) plotted against the overall daily percentage of time spent in that behaviour ($P_D$). The regression lines are solid and the prediction intervals are the dashed lines.
DISCUSSION

Work sampling methods generated rules for obtaining representative samples based on empirically obtained overall relative durations of behaviour. The data analysed were from behaviours with relative duration between 0% and 43.5%. Samples extracted from continuous records were either MTS or continuous (e.g., 100 MTS or 100 s of continuous recording if 100 observations had been prescribed by the equation used). Numbers of MTS observations were dictated by the work sampling equations (Equations 3 and 4), 5-min MTS samples (247 to 327 MTS observations across the week), or 1,000 MTS observations across the week (Rule of 1,000). The main findings were that more representative samples were derived from: 1. more observations, regardless of sampling method; 2. conducting MTS intervals across the day rather than continuous recording for the same number of seconds of observation; and, 3. using the relative accuracy equation (Equation 3) to select the number of observations.

When comparing low- and high-duration behaviours, the equation dictated considerably more observations for lower-duration behaviours. As a result, samples of lower-duration behaviour were more representative. The occurrence of behaviour was not detected in some continuous samples, in some samples extracted using the Rule of 1,000 or 5-min MTS (particularly for low-duration behaviours), and sometimes when the number of momentary time samples was selected using Equation 4. However, the occurrence of behaviour was always detected when the number of momentary time samples was selected using Equation 3 (i.e., when large numbers of samples were taken). The variables that did not affect the representativeness of a sample were: 1. the schedule of MTS observations (fixed, random, or non-continuous); 2. the start time of extracted continuous observation sessions (unless a time nominated by a teacher); and, 3. the source of the initial $P_O$ value (i.e., staff estimate or preliminary observation).
Continuous sampling. The finding that continuous sampling did not produce representative data unless observation sessions are impractically long corroborates the results of Mudford, Beale, and Singh (1990). Although continuous sampling may seem to be preferable to discontinuous sampling because all instances of behaviour are recorded (Johnston & Pennypacker, 2009), the results of Study 1 suggest that continuous sampling should not be considered to be superior to discontinuous sampling in all circumstances. There is a lack of methodological studies using continuous methods, as well as very few studies comparing continuous samples to the full time-of-interest. The results of Study 1 suggest that short (between 10 min and 60 min) continuous observation sessions may not produce representative data. Tiger et al. (2013) found that in their study, increasing the duration of observation sessions increased the representativeness of the sample, but for highly-variable behaviours, fewer than half of the 60-min observations were representative.

The data from Study 1 also suggest the data from continuous observation sessions were poorly representative regardless of the arbitrarily-selected start time of the observation session (i.e., 9 a.m., 11 a.m., 1 p.m., or a randomly-selected start time). However, when teachers were asked to nominate a time during which a particular behaviour of interest could be observed, some of the samples were representative of true values. The timing of the occurrence of behaviours of higher durations (i.e., occurring for ≥ 10% of the week), were more likely to be identified accurately by teachers than lower-duration behaviours (e.g., running for Participant 3, grabbing for Participant 1, and positive affect for Participant 2). The higher-duration behaviours for which representative samples were obtained from teacher-nominated observation times were on-task behaviour (Participants 1 and 2). The daily mean bout duration for on-task behaviour ranged from 139 s to 476 s for Participant 1, and 137 s to 329 s for Participant 2. The daily mean bout duration for lower-duration behaviours were shorter by comparison (e.g., range of 12 s to 21 s for running for Participant
Therefore, teachers were more likely to identify times of occurrence for behaviours that occurred in longer bouts, and therefore were more salient. The mean number of occurrences per day of on-task behaviour for Participants 1 and 2 was 14.6 and 23.6, respectively. For the lower duration behaviours, the mean number of occurrences per day varied greatly (e.g., 14.8 for running for Participant 3, 83.2 for grabbing for Participant 1, and 7.4 for positive affect for Participant 2). Therefore, it appeared that the accuracy with which teachers could identify times of the day during which the behaviour occurred was affected by duration and bout duration rather than the frequency of the behaviour.

On-task behaviour was context-specific (i.e., was affected by the activities in the classroom), perhaps further making the behaviour salient to teachers by being associated with specific tasks (e.g., during sessions using the touch screen white board for Participant 2). Strain and Ezzell (1978) found patterns in the distribution of disruptive and inappropriate behaviours in adolescents in a residential facility. The identification of patterns in the distribution of behaviours allows for interventions to be developed to target settings or activities during which the behaviour is likely. In regards to sampling, the results of Study 1 show that while using informant methods to select observation time may be useful for behaviours that occur in an easily-identifiable context, representativeness is still affected by the sampling method (e.g., short continuous observation sessions produce poorly representative samples).

Of the 14 participant behaviours measured, teachers were only able to nominate specific times of the day for observation sessions for five behaviours (i.e., the teacher responded that "anytime" or "during class time" was a good time to conduct an observation session for the other nine behaviours). Informant methods of gathering information about behaviours have been demonstrated to be less accurate than direct observation. For example, Kazdin, Esveldt-Dawson, and Loar (1983) found only moderate correlations between
checklists completed by teachers and direct observations of children's behaviour in a classroom. They suggested that because teachers have competing duties to fulfil, and are dedicated only to recording behaviour as observers are, their estimates can be less accurate. Therefore, selecting observation session times based on scheduling and convenience is unlikely to affect representativeness because the sampling method has a greater effect on the representativeness of the sample. Informant methods of selecting continuous observation session times may be useful for higher-duration behaviours that are likely to occur in a particular stimulus context, but should not be relied upon for selecting observation session times that will produce representative samples.

**Momentary time sampling using Equation 3.** The finding that more observations conducted with MTS produced more representative data corroborates the results of Brulle and Repp (1984). They found that shorter interval MTS (i.e., shorter intervals between observations and therefore more observations in a session) produced more representative data than MTS with longer intervals (60 s and above). Although the results of Study 1 showed that representative samples were obtained from numbers of MTS selected by Equation 3, large numbers of observations were required (particularly for low-duration behaviours). Obtaining representative samples of low-duration behaviour with small relative error (as dictated by Equation 3) will inevitably require large samples. For many of these extracted samples, the number of observations required per day resulted in very frequent observations (i.e., every 1, 2, or 3 s). Conducting MTS observations this frequently is impractical in applied settings and would require more effort than continuous recording. Therefore, the work sampling Equation 3 may have little utility for determining the number of required observations for low-duration behaviours. Perhaps of more use is the finding that for higher-duration behaviours (i.e., for behaviours that occur for ≥ 30% of the day), more representative
samples were obtained when observations were spread across the day, rather than being conducted continuously.

**Momentary time sampling using Equation 4.** The use of Equation 4 to select the number of required MTS samples produced more representative samples than when continuous observation sessions were conducted, but less representative samples than when Equation 3 was used. Equation 4 sometimes required a very small number of observations (e.g., fewer than five per day), which resulted in large differences between obtained and true values. Equation 4 is likely to produce more representative samples when less stringent accuracy limits are selected, but because less stringent accuracy limits result in changes in what is ‘representative’. For example, Munyisia, Yu, and Hailey (2011) used 50% as the initial $P_O$ value with ± 2% absolute accuracy in Equation 4, resulting in a total of 2,500 observations of how staff were spending their time in a nursing home. The study was conducted over five 8.5-hr days, resulting in one observation every 61 s. In a second nursing home, in which the lowest-duration behaviour had been identified in previous data collection as occurring for 8% of the whole time-of-interest, Munyisia et al. used 8% as the initial $P_O$ value, but with more stringent accuracy limits of ± 1% than those used in Study 1 (± 5%). In total, 2,944 observations were conducted across five days in the second nursing home (588 observations per day, one every 52 s). Although they did not evaluate the representativeness of their samples, it is likely that the larger numbers of observations they selected using Equation 4 produced more representative samples than those produced using Equation 4 in Study 1. Therefore, the use of Equation 4 with more stringent accuracy limits may be of some use in behaviour analysis, if the number of observations is practical. More stringent accuracy requirements will require more observations in both Equations 3 and 4, and desired accuracy must be determined prior to the start of data collection (Tsai, 1996).
Practical numbers of MTS. In applied settings in which a practitioner is under time and resource constraints, and therefore is only able to conduct fewer MTS, the data from Study 1 suggest that higher-duration behaviours will be sampled more representatively than low-duration behaviours. Study 1 evaluated the Rule of 1,000 and 5-min MTS as examples of smaller, more practical numbers of MTS as an alternative to the use of Equations 3 and 4. The results showed that for high-duration behaviours (i.e., behaviours occurring for ≥ 10% of the week), if relative accuracy of approximately ± 20% is tolerable, then samples can be taken using 5-min MTS (the smallest number of observations that were evaluated in Study 1). Kearns, Edwards, and Tingstrom (1990) found that 5- or 10-min MTS could produce sufficiently representative samples, but they only evaluated 1-hr observation sessions and suggested that increases in representativeness would be gained from increasing the duration of observation sessions. Study 1 also evaluated lower-duration behaviours than Kearns et al., who evaluated behaviours occurring for between 20% and 80% of the session. The findings of Study 1 show that 5-min MTS is not likely to produce representative samples for behaviours occurring for < 10% of the week. Therefore, 5-min MTS is not recommended for very low-duration behaviours.

When the Rule of 1,000 was used, more accurate samples were produced than when 5-min MTS was used because the number of observations was greater. However, very low-duration behaviours were not sampled representatively and were likely to be recorded as not occurring when they were (i.e., were missed). In addition, the Rule of 1,000 resulted in MTS intervals likely to be impractical across a full day of observation; MTS interval durations ranged from every 61 s to every 103 s, which is less than one MTS per 2 min.

Work sampling methods usually involve conducting observations across a whole day. In some settings, such as classrooms, observations may only be able to be conducted across part of the day (Doll & Elliott, 1994). Therefore, an evaluation of MTS with different
interval durations across a range of practical observation session durations (i.e., shorter than a full day) would be useful for settings in which MTS cannot be conducted across a full day.

**Schedule of MTS.** The results of Study 1 showed that the method of scheduling MTS observations (i.e., fixed, random, or non-continuous) did not affect the representativeness of the sample. Edwards, Kearns, and Tingstrom (1991) also found no systematic differences between random and fixed MTS of varying interval duration (30 s to 20 min). Edwards et al. suggested that in a classroom setting, a teacher may find it easier to conduct MTS observations at random intervals because the observations can occur when they are convenient to the classroom. However, not only could conducting observations when there are no competing tasks to engage in bias the sample, smaller numbers of observations than desired could be conducted unless the teacher is prompted to record behaviour (e.g., with an audible cue). Scheduling random observations can be more effortful than fixed observations, as a schedule of random times must be generated. Therefore, conducting random samples in applied settings can be impractical (Peregrine, Drews, North, & Slupe, 1993).

Sittig (1993) proposed that fixed-interval sampling is best suited to work (or behavioural) categories that are distributed randomly. Additionally, some studies claim that random sampling, in which the required observations are conducted at randomly-determined times throughout the day, produces less bias than fixed-time sampling (e.g., Oddone & Simel, 1994). However, Finkler et al. (1993) suggested that for behaviours that are random, bias from fixed-time sampling may be negligible, as found for MTS (Powell, Martindale, Kulp, Martindale, & Bauman, 1977). Further analyses of different distributions of behaviour may show the effect of distribution on different sampling methods, and contribute to helping observers choose a sampling method most likely to produce representative data.
**Effort and cost-efficiency.** Sampling methods should produce data sufficiently representative for the purpose of data collection and do so time-efficiently. Direct observation is labour-intensive (Mansell & Beadle-Brown, 2011) and, although none of the observers in the current study reported fatigue, continuous direct observation requires an observer to be dedicated solely to recording behaviour (Gardenier, MacDonald, & Green, 2004). Therefore a compromise between representativeness and cost-efficiency may be necessary. The desired accuracy value for Equations 3 and 4 was selected to be ± 5%, but less stringent accuracy limits may be acceptable in some settings. From a clinical perspective, a sampling method that produces sufficiently representative data is one that detects the smallest clinically-significant behaviour change. Equation 4 required the fewest number of observations but produced the least representative samples, whereas Equation 3 required the most observations and produced the most representative samples. The Rule of 1,000 required fewer observations than Equation 3 but resulted in less representative samples. MTS with 5-min intervals required less effort than both the Rule of 1,000 and Equation 3 (i.e., fewer observations), but produced less representative samples than both. Although conducting 5-min MTS is more likely to be practical than the number of observations required by Equation 3, it could still be impractical for teachers to conduct 5-min MTS across a full school day. In addition, Kearns et al. (1990) found that the most representative samples were produced with shorter-interval MTS (i.e., 30 s), illustrating the need to balance cost with representativeness.

Equation 3 calculates the required numbers of observations based on relative accuracy which results in minute absolute error margins for behaviours occurring for ≤ 10% of the week (e.g., when true duration is 1%, ± 5% absolute error would result in margins of 0.95% and 1.05%). The analyses conducted in Study 1 evaluated both absolute differences from true values (true values subtracted from obtained values) and the number of data points that
fell within the relative accuracy limits. Tiger et al. (2013) suggested that because Mudford et
al. (1990) reported relative error, i.e., error as a percentage of the duration, small errors in
measuring low-duration behaviours were amplified. Tiger et al. (2013) therefore suggested
that reporting absolute errors allows for more equal comparisons of errors for high- and low-
duration behaviours.

However, for low-duration behaviours (e.g., behaviours occurring for < 1% of the
time-of-interest), unrepresentative data may appear representative when absolute accuracy
limits are used. For example, in Study 1, some sessions containing small numbers of
observations produced representative data when the absolute difference between obtained and
true values was calculated. However, a small absolute difference (for example, when a
behaviour is recorded as not occurring, or 0, but occurs during 1% of the time-of-interest,
resulting in an absolute difference of 1%), may not sufficiently reflect true duration even
though the absolute difference would fall on a ± 1% absolute accuracy limit. The view of
Tiger et al. (2013) that acceptable error is likely to be judged according to the purposes of
data collection is logical; there are merits to evaluating both absolute and relative errors.
Perhaps acceptable error varies with overall behavioural duration (e.g., absolute error of ± 5%
may be acceptable for behaviours occurring for 50% of the week, but not for behaviours
occurring for 5% of the week). Mansell (1985) and Alevizos, De Risi, Liberman, Eckman,
and Callahan (1978) suggested that sampling methods that produce representative samples of
higher-duration behaviours may not representatively sample low-duration behaviours.
Therefore, acceptable error (as related to the purpose of data collection) must be considered
when selecting a measurement system for behaviours of varying duration.

Although the use of inferential statistics in behaviour analysis has been deemed
unhelpful by some researchers (Johnston & Pennypacker, 2009), the calibration analyses
classified in Study 1 were a useful method of determining the level of error within which the
obtained duration values fell. As a result, conclusions were able to be drawn regarding the usefulness of the Rule of 1,000 and 5-min MTS if particular levels of accuracy were acceptable. Mudford et al. (2011) demonstrated the usefulness of calibration analyses in behaviour analysis to assess the accuracy and precision of data obtained through direct observation. The calibration analyses conducted in Study 1 further demonstrated the usefulness of such methods in assessing direct observation data.

**Work sampling equations.** In behaviour analysis, on-going and regular inspection of the data, and subsequent changes to interventions, is characteristic of the field (Fahmie & Hanley, 2008), and thus the adjustable work sampling equation might be considered to fit with the regular data inspection in behaviour analysis. Work sampling methods have social validity and are used in other sciences such as nursing research (Pelletier & Duffield, 2003). However, the major limitations of the work sampling equations found in Study 1 were: 1. $P_o$ values that varied greatly across days for the same behaviour resulted in large changes in $N$ (the required number of observations), particularly for Equation 3; 2. for very low-duration behaviours, Equation 3 sometimes dictated a number of observations that equated to one per second or more; and, 3. a lack of detail in the work sampling literature required the inclusion of decision rules regarding the use of the work sampling equations.

In regards to the first limitation, when a $P_o$ value that underestimates the true value of $P_o$ is entered into Equation 3, it will result in the addition of more required observations. In the current study, some of the required numbers of observations were impractically high, even for high-duration behaviours for which Equation 3 dictated the lowest number of observations. As the number of required observations is adjusted based on the current estimate of $P_o$, large changes in the number of required observations would occur concurrently with changes in $P_o$. For behaviours increasing in duration, the number of required observations would decrease, although work sampling texts recommend retaining
the original number planned as this will provide a larger sample size for the same planned effort (e.g., Morris, 1969). However, for decreasing behaviours, the number of required observations increases. Adjusting sampling to increase the likelihood of a representative sample could be desirable as our results show that lower-duration behaviours (i.e., behaviours occurring for \( \leq 10\% \) of the week) are less likely to be sampled representatively with smaller numbers of observations. Increasing the sample size in response to decreasing behavioural duration may strengthen the demonstration of functional control.

In Equation 4, behaviours occurring for 50\% of the time require the largest number of observations, whereas very low- or high-duration behaviours (e.g., behaviours occurring for 10\% or 90\% of the time), require the fewest observations. Changes in the current estimate of \( P_O \) may result in an increase or decrease in the required number of observations selected by Equation 4, depending on the estimate of \( P_O \). However, the differences between required numbers of observations for behaviours of different duration are smaller than differences between required numbers of observations selected by Equation 3. Therefore in practice, changes in the required number of observations are small across changes in \( P_O \) when Equation 4 is used in comparison to the required changes when Equation 3 is used.

When Equation 3 is used to select the required number of observations for decreasing behaviours, the number of required observations increases. Although changes in a measurement system during an intervention may not be desirable traditionally, adjusting sampling to increase the likelihood of a representative sample could be desirable, particularly because the results of Study 1 show that lower-duration behaviours are less likely to be sampled representatively with smaller numbers of observations. Rapp, Colby-Dirksen, Michalski, Carroll, and Lindenberg (2008) found that MTS with smaller intervals (e.g., 10 s) was more likely to detect changes in behavioural duration than MTS with larger intervals, particularly when the changes were large (i.e., a change of 34\% or more in the duration of the
behaviour). The work sampling equation can be used to determine the number of observations required with a desired level of accuracy, but the results of Study 1 demonstrate that the work sampling Equation 3 requires more observations than necessary to produce representative samples when samples are taken using MTS.

In regards to the second limitation, observations that equated to one per second or more result in either continuous sampling (i.e., one observation per second) or impossible-to-conduct numbers of observations when Equation 3 (relative accuracy) is used. The sampling method thus moves away from MTS. Collecting data over a time-of-interest longer than a week could resolve partly the issue of impractically high numbers of observations (as fewer observations could be conducted per day), but would extend the data collection period to impractical or unethical durations (e.g., conducting very long baselines). It is important to note that the current criterion data were collected in a classroom, and that other settings may produce different distributions and durations of behaviour, and different acceptability of sampling methods (and the resulting error).

Poorly described methods and a lack of methodological research using work sampling resulted in a need to develop ad hoc decision rules for Study 1. For example, there does not appear to be any research that guides the selection of a $P_O$ value to enter into the equation when the most recent estimate of $P_O$ is 0, or when $P_O$ is found to be 0 (i.e., no behaviour is detected) in preliminary observations. A lack of reporting of such details in the method results in inconsistencies across studies and does not align with the technological characteristic of applied behaviour analysis, whereby methods are described in sufficient detail to be replicated (Baer, Wolf, & Risley, 1968). Work sampling using continuous observation has been explored theoretically (e.g., Ho & Pape, 2001), and some work sampling studies have used continuous sampling to determine how people spend their time (e.g., Knickman, Lipkin, Finkler, Thompson, & Kiel, 1992). However, there do not appear to
have been any studies using the work sampling equations to select the required number of continuous seconds of observation required to obtain a representative sample. Many of the numbers of required observations selected using Equation 4 were increased to 600 s (10 min) as a minimum time that a practitioner would be likely to spend conducting a direct observation in a classroom. Increasing the number of required continuous seconds of observation to a minimum 600 s negated the used of Equation 4 to select the required number of observations. In conjunction with the finding that samples extracted using numbers of continuous seconds of observation selected by Equation 4 did not produce representative samples, discarding the number of observations selected by Equation 4 leads to the conclusion that using Equation 4 to select the number of continuous seconds of observation is not recommended.

**Initial \( P_0 \) estimates.** Preliminary observations produced more accurate estimates of true duration that teacher estimates, especially for low-duration behaviours. Although there does not appear to be any research on the accuracy of teacher estimates of dimensions of behaviour such as duration or rate, the results corroborate research that has shown informant methods to produce inaccurate information about other aspects of behaviour. For example, Cote, Thompson, Hanley, and McKerchar (2007) found that teacher reports of potential reinforcers for toddlers did not correspond to potential reinforcers identified through a paired-stimulus preference assessment. Although both the items identified by teachers and through direct preference assessments functioned as reinforcers, Cote et al. suggested that a teacher report-informed preference assessment is the most effective method of identifying reinforcers (i.e., superior to teacher report alone).

After reviewing the literature of teacher self-reports of their behaviour in the classroom, Hook and Rosenshine (1979) concluded that teachers do not report their own behaviours accurately. Hook and Rosenshine (1979) suggested that a lack of experience in
comparing estimates to direct observation data contributes to the inaccuracy of teacher estimations. However, in Study 1, teachers estimated the duration of low-duration behaviours less accurately than high-duration behaviours, perhaps because the low-duration behaviours were often challenging behaviours (e.g., grabbing). The duration of challenging low-duration behaviours was overestimated across all teachers. Challenging behaviours may be more salient to teachers because they can be disruptive, causing teachers to overestimate duration. Kazdin et al. (1983) suggested that disruptive behaviours require immediate teacher attention and are therefore more salient than non-disruptive behaviours such as being on-task. Other low-duration behaviours that were overestimated were positive and negative affect. The overestimation of affect may have been a result of unfamiliarity with behavioural indices of affect. Kazdin et al. suggested that behaviours that occur in discrete events (e.g., fighting), are more likely to be observed regardless of the classroom context, whereas behaviours such as on-task behaviour may not be as discrete and therefore not as noticeable to teachers. Indices of affect may fall into the category of less-discrete, and therefore less-salient behaviours.

Although teachers were read the behavioural definitions prior to being asked to estimate behavioural duration, the inaccurate estimates may have resulted from a lack of familiarity with defining overt, observable behaviours. For example, teachers may have estimated the duration of positive affect based on the amount of time they perceived the student to be 'happy', rather than considering the duration of overt indices of positive affect such as smiling. Chafouleas, Jaffery, Riley-Tillman, Christ, and Sen (2013) found that the wording of behavioural definitions affected the accuracy with which observers recorded behaviour (e.g., on-task classroom behaviour was recorded more accurately when defined as on-task behaviour rather than as off-task behaviour). Similarly, Smith, Lambert, and Moore (2013) found that subjective wording in the operational definition of a behaviour (e.g.,
"forcefully") resulted in observers making more errors in recording than an objectively-worded description of the same behaviour. In Study 1, more accurate estimates may have resulted from more thorough explanations of the behaviours of interest (including more examples of inclusions and exclusions), or identifying instances of each behaviour in real time for the teacher prior to recording their estimates of behavioural duration.

Despite the research that cautions the use of teacher reports of behaviour in the classroom, teacher estimates of duration may be sufficient for selecting initial $P_O$ values for work sampling equations because the number of required observations is corrected based on the current estimate of duration. The results of the Study 1 showed no differences in the representativeness of samples collected using teacher estimates and preliminary estimates as the initial $P_O$ value, regardless of the differences in the accuracy of the two types of estimate. Additionally, informant methods such as questionnaires can be more time-efficient than direct observation (Hall, 2005). An advantage of work sampling is incorrect required numbers of observations resulting from inaccurate initial $P_O$ value estimates are corrected over days of observation.

**CONCLUSION**

The empirical approach to representativeness brings into focus the extent to which samples obtained with fewer observations than required may be unrepresentative of the true duration of behaviour. Overall, Study 1 evaluated the work-sampling approach as a way to select objectively the number of MTS observations required to produce representative samples. The use of Equation 3 (relative accuracy) to determine the number of required observations produced more representative samples than Equation 4 (absolute accuracy), but required many more observations. The results demonstrated the relationship between effort and representativeness (i.e., increases in effort result in increases in representativeness). The
work sampling approach was not useful in determining the duration of continuous
observation sessions required for representative samples because continuous observation
sessions produced poorly representative samples.

The use of mathematical models such as Equations 3 and 4 can be advantageous over
more general rules (e.g., shorter MTS intervals are better than long intervals) to choose
measurement systems because they allow more precise selection of representative samples
(Mazur, 2006). When the duration of the behaviour is low (i.e., < 10% of the week),
however, Equation 3 prescribes an unrealistically large number of observations to produce a
desired relative accuracy. It may be that specifying desired absolute accuracy is the only
practical approach in such situations. Equation 3 can be used as an empirical way to
determine when MTS is impractical or when a desired relative accuracy is simply
unobtainable. The practicality of large numbers of observations is likely to be influenced by
the available time in which to conduct observations, which is likely to be determined by
factors such as cost, resources, and the purposes of the data collection. Equation 4 could be
used as an empirical way to predict the likely representativeness of more practical numbers of
observations. Nevertheless, the empirical approach of work sampling brings into focus the
extent to which samples obtained with fewer observations than required may be
unrepresentative of the true behaviour (e.g., those selected by Equation 4).

Further study in the area of metrology in behaviour analysis would be useful to
evaluate the effect of behavioural distribution, and continuous sampling on the
representativeness of data samples. Further research is necessary to develop practical
decision rules that can be referenced by practitioners choosing measurement systems. Such
research would establish the interaction between effort and representativeness of different
sampling durations. The resulting dataset would allow practitioners to select a measurement
system based on the estimated duration of behaviour, desired representativeness (i.e.,
acceptable error), and practicality of the measurement system (i.e., number of required observations). The advantage of such decision rules is the ability to determine empirically how to measure behaviour, an advantage shared by work sampling methods.
The following analysis has been included because it provides data important for the design of Study 2. The current behaviour analytic literature contains no published data on typical ranges of reported durations of behaviour.

Simulations of behavioural events that most closely reflect behaviour observed in applied settings should be informed by empirically-determined ranges of the dimensions of behaviour. In order to determine the range of typical response rates reported in the Journal of Applied Behavior Analysis, Mudford, Locke, and Jeffrey (2011) analysed the range of data points from 60 research articles. They found that 90% of the individual data points represented rates of less than 8.2 responses per minute; however, the overall range of rates was 0 to 104.5 responses per minute. Mudford, Zeleny, Fisher, Krum, and Owen (2011) used this finding to select a range of typical behavior rates to demonstrate the use of calibration in behaviour analysis. They demonstrated the usefulness of calibration analyses in behaviour analysis to assess the accuracy and precision of data obtained through direct observation.

I aimed to further examine the typical ranges of dimensions of behaviour reported by Mudford et al. (2011) by conducting an analysis of the distribution and range of durations reported in the Journal of Applied Behavior Analysis. The primary reason for the analysis was to inform future analyses of the representativeness of samples of simulated behaviours. A second purpose of the analysis was to evaluate the accuracy of a computer-assisted method (digitizing graphical software) of measuring the value of data points.
METHOD

Data selection. Five volumes of the Journal of Applied Behavior Analysis (Vol. 41, 2008 through Vol. 45, 2012) were inspected to identify articles containing data appropriate for inclusion in the study (Appendix 2). Of 397 total empirical articles, 71 were selected for the analysis (17.9%). The selected articles reported either continuous data on the duration of behaviour (i.e., s per session, percentage of session), or discontinuous data on the duration of behaviour (i.e., PIR, WIR, or MTS). A dataset was comprised of all data points from one graph for one participant. Studies were included if the figures contained data from individual direct observation sessions (all but one study were conducted with human participants, the remaining study with a primate), and if duration could be determined by either percentage of intervals, or percentage of session. Seven articles reported absolute duration data in either seconds or minutes; however, percentage of session was able to be calculated from the reported session times. Articles in which duration was reported as total seconds but session length was not reported were excluded as were articles that reported group data or within-session data only.

Measurement. GetData Graph Digitizer 2.26 (getdata-graph-digitizer.com) was used to measure the value of each data point on each figure. A Portable Document Format (pdf) file of each identified article was downloaded from PubMed Central (ncbi.nlm.nih.gov/pubmed) and Microsoft Windows Snipping Tool was used to take a snapshot to create a Joint Photographic Expert Group (jpeg) file for each figure. Each jpeg file was uploaded to the GetData Graph Digitizer 2.26 software. The scale of the Y-axis was set by clicking on the end points and typing in the values, and data points were measured by using the mouse to click on the centre of each point. The software then recorded the value of each data point by measuring them against the scale.

Reliability of measurement. The corresponding authors of one randomly selected article from each issue (19 in total) were contacted. The authors were asked to provide the
spreadsheets from which the figures displaying duration data were made so that the raw data (true values of the data points) could be compared to the data points measured by GetData Graph Digitizer 2.26. The measured data points (y-axis) were plotted against the corresponding true data (x-axis) and conducted a calibration analysis by linear regression in Sigmaplot 11.0 (Systat Software Inc.).

RESULTS

In total, 8,763 data points were measured. Continuous measures of duration were used in 110 datasets (38.6% of a total of 285 datasets; see Table 11). The 110 continuous datasets contained a total of 3,225 data points. The number of articles using continuous measures more than doubled in 2011 and 2012 when compared to previous years. As a result, there were many more datasets containing continuous measures in 2011 and 2012 than in previous years (i.e., 36 and 45 datasets in 2011 and 2012 respectively, compared to seven and three datasets in 2009 and 2010, respectively). There were 175 discontinuous datasets containing 5,538 data points.

Table 11. Number and Percentage of Datasets from Five Volumes of the Journal of Applied Behaviour Analysis in which each Type of Measurement was Used.

<table>
<thead>
<tr>
<th></th>
<th>WIR</th>
<th>PIR</th>
<th>MTS</th>
<th>Continuous</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>0</td>
<td>16</td>
<td>11</td>
<td>45</td>
<td>72</td>
</tr>
<tr>
<td>2011</td>
<td>8</td>
<td>38</td>
<td>24</td>
<td>36</td>
<td>106</td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
<td>22</td>
<td>4</td>
<td>11</td>
<td>39</td>
</tr>
<tr>
<td>2009</td>
<td>0</td>
<td>28</td>
<td>0</td>
<td>7</td>
<td>35</td>
</tr>
<tr>
<td>2008</td>
<td>0</td>
<td>4</td>
<td>18</td>
<td>11</td>
<td>33</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>108</td>
<td>57</td>
<td>110</td>
<td>285</td>
</tr>
<tr>
<td>Percentage</td>
<td>3.5%</td>
<td>37.9%</td>
<td>20%</td>
<td>38.6%</td>
<td></td>
</tr>
</tbody>
</table>
Partial-interval recording was the most commonly used of the discontinuous methods (37.9% of the total number of datasets and 61.7% of the datasets containing discontinuous measures). Whole-interval recording was the least commonly used discontinuous measure, found in 10 datasets (3.6% of the total number of datasets).

For the calibration analysis, five authors responded (26.3%) by providing their data (1,010 data points in total; 11.5% of the total number of data points measured). Reliability was measured by subtracting each raw data point from the corresponding measured data point to calculate the absolute difference between the two. The mean difference was 0.068% (of the session or intervals) with an extreme range of -4.7% to 5.7%, suggesting that the use of GetData Graph Digitizer 2.26 could both under- and overestimate the value of data points (i.e., was not biased in a specific direction).

Figure 25 shows the measured data points plotted against the raw data points. Figure 26 shows an enlarged portion of Figure 25; the measured data points plotted against the raw data points for data points between 40% and 50% (of the session or intervals). The enlarged scale better shows the distribution of data points around the regression line. The results of the calibration analysis by linear regression showed the regression line to fit the data almost perfectly \( R^2 = .9998 \) and that the small standard error of the estimate \( SE = 0.53 \) indicated GetData Graph Digitizer 2.26 to be precise when used to measure data points (Figure 26). The 95% confidence intervals were within \( \pm 0.1\% \). The 95% prediction intervals were within \( \pm 1.0\% \). The narrow confidence and prediction intervals showed the use of GetData Graph Digitizer 2.26 to both produce precise measurements in the current analysis (confidence intervals) and to be predicted to produce precise measurements in the future (precision intervals).
Figure 25. Linear regression analysis comparing data measured with graph digitizing software to original data obtained from article authors. The regression line is solid and the dashed lines represent the 95% prediction intervals.

Figure 26. Enlarged portion of Figure 25 showing linear regression analysis comparing data measured with graph digitizing software to original data obtained from article authors for data points between 40% and 50% of the session or intervals. The regression line is solid and the dashed lines represent the 95% prediction intervals.
DISTRIBUTION OF BEHAVIOURAL DURATIONS

Measured data points were separated by continuous recording (percentage of session) and discontinuous recording (percentage of intervals) and grouped into bins. All data points were grouped into bins of 10% (e.g., > 10% to < 20%, > 20% to < 30%). Data points of 0% and 100% were grouped into two separate bins. Figure 27 shows the cumulative frequency of data points across increasing duration (percentage of session and percentage of intervals).

For both continuous and discontinuous measures, the minimum duration was 0% and the maximum was 100%. The modal duration was 0% (11.8% of all continuous data points and 20.5% of all discontinuous data points). Figure 27 shows that 34.5% of continuous data points and 39.6% of discontinuous data points fell between 0% and 10%. Of continuous and discontinuous data points (respectively), 3.1% and 3.8% were 100% (of the session or intervals). The solid and broken lines on Figure 27 show that 50% of data points were interpolated as less than 30% of the session for continuous measures and 20% of intervals for discontinuous measures (respectively).

Figure 27. Cumulative number of data points recorded in bands of 10% (of intervals for discontinuous measures and of the session for continuous measures, dotted line with open circles and solid line with filled circles respectively).
Figure 28 shows the range of values for each dataset. The range included 0% of the session or intervals in 50.9% and 62.9% of the continuous and discontinuous datasets respectively. Of the continuous and discontinuous datasets (respectively), 73.6% and 84% contained values of 50% or more of the session or intervals.

*Figure 28.* Maximum and minimum values reported in each data set (i.e., range). The top of each grey bar shows the maximum value and the bottom of each grey bar shows the minimum value. Datasets are ordered from left to right by the maximum measured duration in each dataset.
DISCUSSION

The data from 71 articles from the Journal of Applied Behavior Analysis were evaluated to determine the range of reported values of continuous and discontinuous measures of duration (percentage of session and percentage of intervals, respectively). First, I found that although the range of data points was broad (0% to 100%), more than half of the data points were between 0% and 30% (percentage of session or intervals). Second, a greater number of articles reported data collected by discontinuous measures than continuous measures of duration, although the distribution of data points across the range of duration values was similar between continuous and discontinuous measures. Third, the nonoccurrence of a behaviour (0%) was reported more frequently when discontinuous rather than continuous methods were used. Fourth, GetData Graph Digitizer 2.26 was found to obtain accurately the values of data points from published graphs.

The finding that the distribution of reported behavioural durations is skewed towards behaviours that occur for 30% of the time or less is similar to the findings of Mudford, Locke, et al (2011). They reported a skewed distribution towards response rates of less than 1.0 response per minute. However, the ranges of values in the current duration datasets were generally larger than those found in the rate datasets measured by Mudford, Locke, et al. (2011). The implications for studies simulating behavioural duration are that a range of durations from 0% to 100% should be simulated; however that extra consideration should be given to durations less than 30%, as these represent a large proportion of individual sessions of recording behavioural duration. For example, Wirth et al. (2014) simulated behavioural durations in 1% increments from 1% to 100%; however, Powell, Martindale, Kulp, Martindale, and Bauman (1977) contrived higher-duration behaviours; 20%, 50%, and 100% of the session.

Although the data produced by Powell et al. (1977) informs the use of MTS with high-duration behaviours, there is little research to guide the sampling of very low-duration
behaviours. Wirth et al. (2014) found that, despite previous reports that MTS does not produce any systematic error, MTS was more likely to underestimate duration at low behavioural-durations. Similarly, Gardenier, MacDonald, and Green (2004) found that MTS with 10-s intervals produced reasonably representative samples of behaviour across overall durations ranging from approximately 5% to 55%, but that error was larger for low-duration behaviours. Given the large range of durations found within datasets, choosing a measurement system that samples both lower- and higher-duration behaviours with minimal error may be important. More conservative measurement systems are recommended when the goal of a behaviour change program is to increase the duration of a behaviour (Mayer, Sulzer-Azaroff, & Wallace, 2012). Successful programs to increase or decrease behavioural duration are likely to produce datasets with large ranges. Therefore, it may be important to select a measurement method likely to detect changes in behaviour across a range of durations. For example, Meany-Daboul, Roscoe, Bourret, and Ahearn (2007) found that when measuring behaviours of decreasing duration, a functional relation may go undetected by MTS or PIR if the changes in duration are small. The purposes of the data collection and the size of the desired change in duration may affect the choice of measurement method.

Further research on representative measurement systems for behaviours of changing duration is required, particularly given the large range of durations found within some of the datasets sampled in our study.

Mudford, Taylor, and Martin (2009) reviewed volumes of the *Journal of Applied Behavior Analysis* from 1995 to 2005, finding that 55% of the articles used continuous measures of human operant behaviour. Of the articles using continuous methods, 36% reported duration measures and 95% used frequency measures. Although the current data are not directly comparable as continuous measures of frequency were not included, the current data do suggest the continuous measures of duration are being used increasingly more often,
DISTRIBUTION OF BEHAVIOURAL DURATIONS

corroborating one finding of Mudford et al. (2009). However, despite the evidence that MTS
does not produce systematic biases in estimating duration and is thus recommended over PIR
and WIR (e.g., Harrop & Daniels, 1986; Meany-Daboul et al., 2007), the current analysis
found PIR to be the most commonly reported discontinuous measure of duration.

Although the distribution of duration values was similar across discontinuous and
continuous measures, 54.6% and 56.1% of the data points were less than 30% of the session
(continuous measures) or intervals (discontinuous measures). The use of PIR in 61.7% of
datasets using discontinuous measures may account for the difference between the two
measures, because PIR overestimates duration (Powell et al., 1977). Small sample size from
a single journal, as identified as a potential issue by Mudford, Locke, et al. (2011), or
differences in the settings or behaviours for which discontinuous or continuous measures are
likely to be selected could have affected the results.

Choosing a measurement system is likely to involve consideration of the purposes of
data collection (Taylor, Skourides, & Alvero, 2012) and the constraints of the setting. For
example, MTS may be selected as it does not require the observer to observe throughout the
interval, unlike both PIR and WIR (Quera, 1990). Further analysis of any differences
between continuous and discontinuous measures may be warranted as, although continuous
measures are being used increasingly, discontinuous measures are still used and may be the
method of choice in some settings. For example, Rapp et al. (2007) suggested that MTS may
be preferred over continuous recording in some settings as the observer does not need to
record onsets and offsets of the behaviour. As discontinuous measures continue to be used, it
is important to know whether discontinuous measures are able to detect changes in duration
(as represented in many of the datasets in the current study). In the first of a series of studies
evaluating the utility of discontinuous measures in detecting changes in duration, Rapp et al.
(2008) showed that MTS with 10-s intervals was superior to PIR in detecting changes in
duration, and that planned large changes in duration (e.g., from 75% of the session to 0% of the session) can be measured acceptably with MTS with larger intervals (e.g., 30 s). Rapp et al. (2008) suggested that conducting an initial 10-min baseline session using continuous recording will indicate the direction of desired change and current duration and thus guide the choice of appropriate discontinuous measures.

The finding that more discontinuous than continuous data points were reported to be 0% (nonoccurrence of a behaviour) cannot be explained easily. Data were recorded on the minimum and maximum session lengths and were reviewed as a possible explanation. As longer sessions (e.g., 60 min) have been shown to produce more representative data than shorter sessions (Tiger et al., 2013), shorter sessions may suggest the likelihood of false negatives (i.e., failing to identify the occurrence of a behaviour). However, the current analysis found no difference between the range and distribution of session lengths between studies using continuous and discontinuous methods. Decreases in representativeness with increases in interval length in discontinuous recording have been reported (e.g., Harrop & Daniels, 1986), however 73.5% of the articles using discontinuous measures used intervals of 10 s or less. Similarly, the current analysis found no differences between the proportions of continuous and discontinuous methods measuring increases or decreases in the behaviour of interest. None of the measured variables appear to offer an explanation for the larger proportion of 0% data points obtained with discontinuous measures. Further investigation either through the analysis of a larger number of articles published in other journals or more thorough analyses of the possible contributing variables is required.

A secondary purpose of the current study was to evaluate the utility of the GetData Graph Digitizer 2.26 software. The software was found to be easy to use, and measured data points accurately. The only potential difficulty is in the symbol of the data points. It was noted that it was more difficult to identify and click on the centre of a triangle data point than
a circle, resulting in some human error. However, the zoom function was of use to overcome this difficulty and the results of the calibration by linear regression showed errors to be minimal. Although there are few published studies using graph digitizing software and none that I know of using GetData Graph Digitizer, other behavioural researchers have shown interest in the tool. For example, Parker, Hagan-Burke, and Vannest (2007) used i-extractor (Linden Software, 1998) to assess a method of determining effect size. To assess the accuracy of the tool, they graphed the values obtained with the software and after resizing the graphs to match the originals, used visual inspection to compare the two. Parker et al. reported that comparing the graphs visually allowed human errors in the obtained values to be identified, however no further analyses of the software were conducted. I suggest that further research into the uses of graph digitizing software is timely.

In order to further inform methodological studies in behavioural measurement, further evaluation of reported dimensions of behaviour such as bout duration and inter-response time is required. Despite a paucity of published research focused on dimensions of behaviour such as bout duration, pooling data from a range of researchers in various settings may allow for the development of typical ranges of parameters for use in simulation studies. The data from the current analysis complement the literature on behavioural measurement and provide a preliminary evaluation of graph digitizing software.
The representativeness of sampling methods is affected by the parameters of the measurement system (e.g., interval duration) and the dimensions of behaviour (e.g., duration, bout duration), and thus the interaction between dimensions of behaviour and parameters of measurement systems merits further study. Powell (1984a) suggested that for partial interval recording, a lack of prior knowledge of the dimensions of behaviour (such as duration) and arbitrarily-selected observation session durations will result in unrepresentative samples. The same is likely to be true of momentary time sampling. The behavioural literature on representativeness is able to provide general rules for selecting representative measurement systems (e.g., using shorter intervals between observations will, in general, produce more representative samples). For example, Hartmann and Wood (1982) described the components of a measurement system such as selecting the medium of recording, the schedule of observation systems, and the sampling procedure, and described the options available to practitioners for each component. General rules help practitioners to choose measurement systems based on known advantages and limitations to particular recording methods. However, applied behaviour analysis would benefit from an empirical method of choosing a measurement system based on the known or estimated dimensions of behaviour, acceptable error, and the parameters of a measurement system likely to be practical in the setting.

Rojahn and Kanoy (1985) used simulated behavioural events of varied frequency, duration, and distribution using Monte Carlo methods to evaluate the interaction between
dimensions of behaviour and measurement system parameters on the representativeness of
the obtained samples. Their simulated time sampling procedure involved an observation
interval followed by an interval in which the simulated observer would record the presence or
absence of the behaviour. Rojahn and Kanoy varied both the absolute duration of the
observation intervals and the ratio of observation interval duration to recording interval
duration. They developed tables to show the predicted level of error when a measurement
system was selected based on estimated or known values of the dimensions of behaviour (i.e.,
rate, duration, and distribution). Rojahn and Kanoy suggested that the development of tables
or figures to which practitioners can refer as a data-based method of selecting a measurement
system would be useful, because general rules (e.g., Hartmann & Wood, 1982) are not
prescriptive enough (i.e., are too general or complex to be practical). Rojahn and Kanoy
acknowledged that their data represented only a small range of values for each of the
parameters, and that a data-based tool for selecting a measurement system would not account
for observer errors. However, some observer errors can be related to the sampling method
(e.g., observers make more errors using PIR than MTS; Murphy & Harrop, 1994), and
therefore may be considered when selecting a measurement method.

Despite the attempts of some studies to use mathematical models to predict the
representativeness of samples of behaviour, the results are complex and the practical
implications unclear. For example, Rogosa and Ghandour (1991) provided mathematical
models to compare the effect of factors such as conducting one long observation session
rather than multiple short sessions, and varying observation interval duration. Although
acknowledged as an exploratory analysis into the use of mathematical models in the direct
observation of behaviour, the data presented by Rogosa and Ghandour were complex and
include statistics more advanced than most behaviour analytic practitioners would have
encountered. In addition, standard statistical models do not include constructs specific to
behaviour analysis (such as response rate or duration), therefore developing mathematical models specific to behaviour may be of greater use (Fisher & Lerman, 2014).

In order to evaluate the effect of both varying the dimensions of behaviour and using different measurement systems on the representativeness of samples of behaviour in a more practical way, some studies have used simulations. For example, Green and Alverson (1978) developed a computer program to simulate behaviours of varying duration and inter-response time. Their data showed MTS to produce less error than PIR and WIR, and that interval duration and behavioural duration both affected the degree of error produced. In a more recent example, Wirth, Slaven, and Taylor (2014) developed a computer program to simulate behaviours of varying bout duration and overall duration, as well as simulating MTS, PIR, and WIR of the behaviours. Their results corroborated those of previous studies regarding the systematic errors of the three discontinuous recording methods, and that shorter intervals result in more representative data. Wirth et al. suggested that although simulations may not account for variations in the distribution of behaviours in applied settings and may lack generality, they are a useful method for studying the effect of controlled variations in dimensions of behaviour and sampling methods due to the control over the variable they permit and the elimination of human observation errors.

The first aim of Study 2 was to use computer simulations to evaluate the effect of varying duration and bout duration on the representativeness of samples extracted using simulated MTS. Bout durations and overall durations were selected empirically from the data collected in Study 1 and the review of reported measures of duration in the *Journal of Applied Behavior Analysis*. From the data generated and analysed, the second aim of Study 2 was to develop a set of decision rules from which practitioners are able to select a measurement system based on estimates of the values of dimensions of the behaviour (bout
duration and overall duration), and the acceptable error. Finally, the decision rules were applied to the whole-week datasets obtained in Study 1 to test their use with naturalistic data.

**METHOD**

**Procedure**

Simulations of behaviours and extracted samples were conducted using MATLAB® R2011b (Mathworks). Appendix 3 shows a sample of the code used to simulate behaviours and sampling methods. Eight-hr streams of behaviours of low (1%, 2%, and 5%), intermediate (10% and 20%), and high overall duration (50%, 75%, and 90%) with varied bout duration (1 s to 10 s, 30 s to 90 s 120 s to 300 s, and 600 s to 1,200 s) were simulated. Each combination of overall duration and bout duration was simulated 100 times (e.g., behaviour of 10% overall duration with short bout duration). From each simulated 8-hr stream of behaviour, MTS samples with intervals of 10 s, 15 s, 20 s, 30 s, 60 s, 300 s, 600 s, 1,800 s, and 3,600 s were extracted across session durations ranging from 10 min to 480 min (8 hr). The resultant number of data points was 100 for each combination of MTS interval and session duration (e.g., 10-s MTS across a 30-min session) for each combination of overall and bout duration.

**Simulated behaviours.** Overall durations of 1%, 2%, 5%, 10%, 20%, 50%, 75%, and 90% were selected based on the range of durations reported in the review of five years of the *Journal of Applied Behavior Analysis*. The daily mean bout duration of each behaviour from Study 1 was obtained from the summary statistics function in ObsWin32 (antam.co.uk). Mean bout durations were divided into 10-s bins and the percentage of behaviours with mean bout durations falling within each bin (e.g., between 1 s and 10 s, or 11 s and 20 s) was calculated (Figure 29).
Figure 29. Mean bout durations from Study 1 (in 10-s bins) and the percentage of behaviours with mean bout durations in each 10-s bin. The short, medium, and long bout ranges selected for Study 2 are indicated by the dashed boxes; 1 s to 10 s, 30 s to 60 s, and 120 s to 300 s.

Based on the distribution of mean bout durations, the bout ranges selected for simulations were 1 s to 10 s, 30 s to 60 s, and 120 s to 300 s (labelled short, medium, and long bouts, respectively). The data from Study 1 showed that some occurrences of behaviour were in long bouts (e.g., one occurrence of unhappy was recorded as occurring for 1,294 s for Participant 1), and not reflected in the means due to the large range of bout durations. Therefore, for behaviours occurring for $\geq 10\%$ of the day, a very long bout range was also simulated (600 s to 1,200 s). Very long bout ranges were not simulated for behaviours occurring for $< 1\%$ of the day because the total required seconds of behaviour was less than the minimum value of the very long bout range (600 s). Short bout ranges were not simulated for behaviour occurring for 90% of the day as the total number of seconds in which the behaviour was not occurring (2,880 s, 10% of the day) was less than the required number of interbout intervals (5,184 required events, and 5,185 required interbout intervals). Therefore,
the bouts were unable to be separated by the specified minimum of 1 s. Each overall duration was paired with each bout duration (28 combinations in total) (Table 12).

Each simulated day was 8 hr (28,800 s) in duration. To simulate a behaviour occurring for an overall 10% of the day, a total of 2,880 s of behaviour was required. In order to determine the number of behavioural events (i.e., occurrences) to be simulated ($N_E$), the total number of required time (Total$_s$) was divided by the midpoint of the bout range (Equation 7). $B_{max}$ represents the maximum bout duration (e.g., 90 s) and $B_{min}$ represents the minimum bout duration (e.g., 30 s) for a given bout duration range (e.g., medium bout range).

$$N_E = \frac{\text{Total}_s}{0.5 (B_{max} - B_{min}) + B_{min}}$$  \hspace{1cm} (7)

The duration of each interbout interval was selected randomly by the program but was required to be a minimum of 1 s. Table 12 shows the number of events required for each combination of overall duration and bout duration.

Part of the code used to simulate behaviours was obtained from an online file exchange for users of MATLAB® (Stafford, 2006), and help with programming was sought in an online MATLAB® user community when required (Bobrov, 2013; Roberson, 2013; Stafford, 2013, 2014).
Table 12. Number of Events Simulated for each Combination of Overall Duration and Bout Duration Range Derived from Equation 7.

<table>
<thead>
<tr>
<th>Overall duration</th>
<th>Total s</th>
<th>1 – 10 s</th>
<th>30 – 90 s</th>
<th>120 – 300 s</th>
<th>600 – 1,200 s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>288 s</td>
<td>58</td>
<td>5</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>2%</td>
<td>576 s</td>
<td>115</td>
<td>10</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>5%</td>
<td>1,440 s</td>
<td>288</td>
<td>24</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>10%</td>
<td>2,880 s</td>
<td>576</td>
<td>48</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>20%</td>
<td>5,760 s</td>
<td>1,152</td>
<td>96</td>
<td>27</td>
<td>6</td>
</tr>
<tr>
<td>50%</td>
<td>14,400 s</td>
<td>2,880</td>
<td>240</td>
<td>69</td>
<td>16</td>
</tr>
<tr>
<td>75%</td>
<td>21,600 s</td>
<td>4,320</td>
<td>360</td>
<td>103</td>
<td>24</td>
</tr>
<tr>
<td>90%</td>
<td>25,920 s</td>
<td>-</td>
<td>432</td>
<td>123</td>
<td>29</td>
</tr>
</tbody>
</table>

Momentary time sampling (fixed-interval). MTS with interval durations of 10 s, 15 s, 20 s, 30 s, 60 s, 300 s, 600 s, 1,800 s, and 3,600 s was simulated. A range of observation session durations from 10 min to 480 min (8 hr) in 10-min increments was simulated. The program code identified the whether behaviour was occurring at each MTS observation. The percentage of observations in which the behaviour was occurring was calculated ($P_O$, the number of observations in which the behaviour was occurring divided by the total number of observations in the session and multiplied by 100), and compared with the true percentage of time the behaviour was occurring across the 8-hr day. Table 13 shows the number of MTS observations conducted for each combination of MTS interval and session duration.
Table 13.
*Number of Observations Conducted Across an 8-hr Day for the Range of MTS Interval Durations and Session Durations.*

<table>
<thead>
<tr>
<th>Session duration</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>30</th>
<th>60</th>
<th>300</th>
<th>600</th>
<th>1,800</th>
<th>3,600</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 min</td>
<td>60</td>
<td>40</td>
<td>30</td>
<td>20</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>20 min</td>
<td>120</td>
<td>80</td>
<td>60</td>
<td>40</td>
<td>20</td>
<td>4</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>30 min</td>
<td>180</td>
<td>120</td>
<td>90</td>
<td>60</td>
<td>30</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>40 min</td>
<td>240</td>
<td>160</td>
<td>120</td>
<td>80</td>
<td>40</td>
<td>8</td>
<td>4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50 min</td>
<td>300</td>
<td>200</td>
<td>150</td>
<td>100</td>
<td>50</td>
<td>10</td>
<td>5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1 hr 60 min</td>
<td>360</td>
<td>240</td>
<td>180</td>
<td>120</td>
<td>60</td>
<td>12</td>
<td>6</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>70 min</td>
<td>420</td>
<td>280</td>
<td>210</td>
<td>140</td>
<td>70</td>
<td>14</td>
<td>7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>80 min</td>
<td>480</td>
<td>320</td>
<td>240</td>
<td>160</td>
<td>80</td>
<td>16</td>
<td>8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>90 min</td>
<td>540</td>
<td>360</td>
<td>270</td>
<td>180</td>
<td>90</td>
<td>18</td>
<td>9</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>100 min</td>
<td>600</td>
<td>400</td>
<td>300</td>
<td>200</td>
<td>100</td>
<td>20</td>
<td>10</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>110 min</td>
<td>660</td>
<td>440</td>
<td>330</td>
<td>220</td>
<td>110</td>
<td>22</td>
<td>11</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2 hr 120 min</td>
<td>720</td>
<td>480</td>
<td>360</td>
<td>240</td>
<td>120</td>
<td>24</td>
<td>12</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>130 min</td>
<td>780</td>
<td>520</td>
<td>390</td>
<td>260</td>
<td>130</td>
<td>26</td>
<td>13</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>140 min</td>
<td>840</td>
<td>560</td>
<td>420</td>
<td>280</td>
<td>140</td>
<td>28</td>
<td>14</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>150 min</td>
<td>900</td>
<td>600</td>
<td>450</td>
<td>300</td>
<td>150</td>
<td>30</td>
<td>15</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>160 min</td>
<td>960</td>
<td>640</td>
<td>480</td>
<td>320</td>
<td>160</td>
<td>32</td>
<td>16</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>170 min</td>
<td>1,020</td>
<td>680</td>
<td>510</td>
<td>340</td>
<td>170</td>
<td>34</td>
<td>17</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3 hr 180 min</td>
<td>1,080</td>
<td>720</td>
<td>540</td>
<td>360</td>
<td>180</td>
<td>36</td>
<td>18</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>190 min</td>
<td>1,140</td>
<td>760</td>
<td>570</td>
<td>380</td>
<td>190</td>
<td>38</td>
<td>19</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200 min</td>
<td>1,200</td>
<td>800</td>
<td>600</td>
<td>400</td>
<td>200</td>
<td>40</td>
<td>20</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>210 min</td>
<td>1,260</td>
<td>840</td>
<td>630</td>
<td>420</td>
<td>210</td>
<td>42</td>
<td>21</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>220 min</td>
<td>1,320</td>
<td>880</td>
<td>660</td>
<td>440</td>
<td>220</td>
<td>44</td>
<td>22</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Time</td>
<td>Value 1</td>
<td>Value 2</td>
<td>Value 3</td>
<td>Value 4</td>
<td>Value 5</td>
<td>Value 6</td>
<td>Value 7</td>
<td>Value 8</td>
<td>Value 9</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>230 min</td>
<td>1,380</td>
<td>920</td>
<td>690</td>
<td>460</td>
<td>230</td>
<td>46</td>
<td>23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 hr</td>
<td>240 min</td>
<td>1,440</td>
<td>960</td>
<td>720</td>
<td>480</td>
<td>240</td>
<td>48</td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>250 min</td>
<td>1,500</td>
<td>1,000</td>
<td>750</td>
<td>500</td>
<td>250</td>
<td>50</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>260 min</td>
<td>1,560</td>
<td>1,040</td>
<td>780</td>
<td>520</td>
<td>260</td>
<td>52</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>270 min</td>
<td>1,620</td>
<td>1,080</td>
<td>810</td>
<td>540</td>
<td>270</td>
<td>54</td>
<td>27</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>280 min</td>
<td>1,680</td>
<td>1,120</td>
<td>840</td>
<td>560</td>
<td>280</td>
<td>56</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>290 min</td>
<td>1,740</td>
<td>1,160</td>
<td>870</td>
<td>580</td>
<td>290</td>
<td>58</td>
<td>29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 hr</td>
<td>300 min</td>
<td>1,800</td>
<td>1,200</td>
<td>900</td>
<td>600</td>
<td>300</td>
<td>60</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>310 min</td>
<td>1,860</td>
<td>1,240</td>
<td>930</td>
<td>620</td>
<td>310</td>
<td>62</td>
<td>31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>320 min</td>
<td>1,920</td>
<td>1,280</td>
<td>960</td>
<td>640</td>
<td>320</td>
<td>64</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>330 min</td>
<td>1,980</td>
<td>1,320</td>
<td>990</td>
<td>660</td>
<td>330</td>
<td>66</td>
<td>33</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>340 min</td>
<td>2,040</td>
<td>1,360</td>
<td>1,020</td>
<td>680</td>
<td>340</td>
<td>68</td>
<td>34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>350 min</td>
<td>2,100</td>
<td>1,400</td>
<td>1,050</td>
<td>700</td>
<td>350</td>
<td>70</td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 hr</td>
<td>360 min</td>
<td>2,160</td>
<td>1,440</td>
<td>1,080</td>
<td>720</td>
<td>360</td>
<td>72</td>
<td>36</td>
<td>12</td>
</tr>
<tr>
<td>370 min</td>
<td>2,220</td>
<td>1,480</td>
<td>1,110</td>
<td>740</td>
<td>370</td>
<td>74</td>
<td>37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>380 min</td>
<td>2,280</td>
<td>1,520</td>
<td>1,140</td>
<td>760</td>
<td>380</td>
<td>76</td>
<td>38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>390 min</td>
<td>2,340</td>
<td>1,560</td>
<td>1,170</td>
<td>780</td>
<td>390</td>
<td>78</td>
<td>39</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>400 min</td>
<td>2,400</td>
<td>1,600</td>
<td>1,200</td>
<td>800</td>
<td>400</td>
<td>80</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>410 min</td>
<td>2,460</td>
<td>1,640</td>
<td>1,230</td>
<td>820</td>
<td>410</td>
<td>82</td>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 hr</td>
<td>420 min</td>
<td>2,520</td>
<td>1,680</td>
<td>1,260</td>
<td>840</td>
<td>420</td>
<td>84</td>
<td>42</td>
<td>14</td>
</tr>
<tr>
<td>430 min</td>
<td>2,580</td>
<td>1,720</td>
<td>1,290</td>
<td>860</td>
<td>430</td>
<td>86</td>
<td>43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>440 min</td>
<td>2,640</td>
<td>1,760</td>
<td>1,320</td>
<td>880</td>
<td>440</td>
<td>88</td>
<td>44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>450 min</td>
<td>2,700</td>
<td>1,800</td>
<td>1,350</td>
<td>900</td>
<td>450</td>
<td>90</td>
<td>45</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>460 min</td>
<td>2,760</td>
<td>1,840</td>
<td>1,380</td>
<td>920</td>
<td>460</td>
<td>92</td>
<td>46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>470 min</td>
<td>2,820</td>
<td>1,880</td>
<td>1,410</td>
<td>940</td>
<td>470</td>
<td>94</td>
<td>47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 hr</td>
<td>480 min</td>
<td>2,880</td>
<td>1,920</td>
<td>1,440</td>
<td>960</td>
<td>480</td>
<td>96</td>
<td>48</td>
<td>16</td>
</tr>
</tbody>
</table>

178
Assumptions

Somewhat arbitrary assumptions regarding acceptable session duration and accuracy limits were selected for the purpose of data analysis, although these may not be acceptable in some settings or for some behaviours. A session duration of 60 min was the maximum duration assumed to be practical in applied settings.

Accuracy limits. The relative (Table 14) and absolute (Table 15) accuracy limits of ± 1% to ± 20% for behaviours occurring for a range of durations were calculated. For example, relative accuracy limits of ± 5% for a behaviour with 1% overall duration indicate that data points should fall within 0.95% and 1.05%. Absolute accuracy limits of ± 5% for a behaviour with 1% overall duration indicate that data points should fall within 0% and 6%. Figure 30 shows relative and absolute accuracy limits of 5%, 10%, 15%, and 20% for behaviours of overall duration ranging from 1% to 100%. The dashed diagonal line represents perfect correspondence between obtained and true values. Figure 30 shows that relative accuracy limits were more stringent than absolute accuracy limits, particularly for lower-duration behaviours. It was assumed (through visual inspection of Table 14) that any relative accuracy limits that were likely to be desirable were captured by using absolute accuracy limits. For example, for a behaviour occurring with 10% overall duration, relative accuracy limits of ±20% (data will fall within 8% and 12%) are the same as absolute accuracy limits of ±2%. In addition, absolute accuracy limits are easier to calculate. Therefore, relative accuracy limits were removed from the data analysis.
Figure 30. Absolute (left panel) and relative (right panel) accuracy limits for behaviours occurring for overall durations from 0% to 100% of the time-of-interest. Perfect correspondence between obtained and true values is represented by the dashed line.

For low-duration behaviours (1%, 2%, and 5% overall duration) absolute accuracy limits of up to ±5% were assumed to be acceptable in applied settings. For intermediate- and high-duration behaviours (between 10% and 90% overall duration), absolute accuracy limits of up to ±10% were assumed to be acceptable in applied settings. Acceptable absolute accuracy limits are indicated in the grey boxes in Table 15.
Table 14.
Upper and Lower Relative Accuracy Limits of ±1% to ±20% Calculated for Overall Durations Ranging from 1% to 90%.

<table>
<thead>
<tr>
<th>Accuracy limits</th>
<th>1%</th>
<th>2%</th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>50%</th>
<th>75%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>± 1%</td>
<td>0.99% - 1.98%</td>
<td>4.95% - 9.9%</td>
<td>19.8% - 49.5%</td>
<td>74.25% - 89.1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 2%</td>
<td>0.98% - 1.96%</td>
<td>4.9% - 9.8%</td>
<td>19.6% - 49%</td>
<td>73.5% - 88.2%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 3%</td>
<td>0.97% - 1.94%</td>
<td>4.85% - 9.7%</td>
<td>19.4% - 48.5%</td>
<td>72.75% - 87.3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 4%</td>
<td>0.96% - 1.92%</td>
<td>4.8% - 9.6%</td>
<td>19.2% - 48%</td>
<td>72% - 86.4%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 5%</td>
<td>0.95% - 1.9%</td>
<td>4.75% - 9.5%</td>
<td>19% - 47.5%</td>
<td>71.25% - 85.5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 6%</td>
<td>0.94% - 1.88%</td>
<td>4.7% - 9.4%</td>
<td>18.8% - 47%</td>
<td>70.5% - 84.6%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 7%</td>
<td>0.93% - 1.86%</td>
<td>4.65% - 9.3%</td>
<td>18.6% - 46.5%</td>
<td>69.75% - 83.7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 8%</td>
<td>0.92% - 1.84%</td>
<td>4.6% - 9.2%</td>
<td>18.4% - 46%</td>
<td>69% - 82.8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 9%</td>
<td>0.91% - 1.82%</td>
<td>4.55% - 9.1%</td>
<td>18.2% - 45.5%</td>
<td>68.25% - 81.9%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 10%</td>
<td>0.9% -</td>
<td>1.8% -</td>
<td>4.5% -</td>
<td>9% -</td>
<td>18% -</td>
<td>45% -</td>
<td>67.5% -</td>
<td>81% -</td>
</tr>
<tr>
<td>± 11%</td>
<td>0.89% -</td>
<td>1.78% -</td>
<td>4.45% -</td>
<td>8.9% -</td>
<td>17.8% -</td>
<td>44.5% -</td>
<td>66.75% -</td>
<td>80.1% -</td>
</tr>
<tr>
<td>± 12%</td>
<td>0.88% -</td>
<td>1.76% -</td>
<td>4.4% -</td>
<td>8.8% -</td>
<td>17.6% -</td>
<td>44% -</td>
<td>66% -</td>
<td>79.2% -</td>
</tr>
<tr>
<td>± 13%</td>
<td>0.87% -</td>
<td>1.74% -</td>
<td>4.35% -</td>
<td>8.7% -</td>
<td>17.4% -</td>
<td>43.5% -</td>
<td>65.25% -</td>
<td>78.3% -</td>
</tr>
<tr>
<td>± 14%</td>
<td>0.86% -</td>
<td>1.72% -</td>
<td>4.3% -</td>
<td>8.6% -</td>
<td>17.2% -</td>
<td>43% -</td>
<td>64.5% -</td>
<td>77.4% -</td>
</tr>
<tr>
<td>± 15%</td>
<td>0.85% -</td>
<td>1.7% -</td>
<td>4.25% -</td>
<td>8.5% -</td>
<td>17% -</td>
<td>42.5% -</td>
<td>63.75% -</td>
<td>76.5% -</td>
</tr>
<tr>
<td>± 16%</td>
<td>0.84% -</td>
<td>1.68% -</td>
<td>4.2% -</td>
<td>8.4% -</td>
<td>16.8% -</td>
<td>42% -</td>
<td>63% -</td>
<td>75.6% -</td>
</tr>
<tr>
<td>± 17%</td>
<td>0.83% -</td>
<td>1.66% -</td>
<td>4.15% -</td>
<td>8.3% -</td>
<td>16.6% -</td>
<td>41.5% -</td>
<td>62.25% -</td>
<td>74.7% -</td>
</tr>
<tr>
<td>± 18%</td>
<td>0.82% -</td>
<td>1.64% -</td>
<td>4.1% -</td>
<td>8.2% -</td>
<td>16.4% -</td>
<td>41% -</td>
<td>61.5% -</td>
<td>73.8% -</td>
</tr>
<tr>
<td>± 19%</td>
<td>0.81% -</td>
<td>1.62% -</td>
<td>4.05% -</td>
<td>8.1% -</td>
<td>16.2% -</td>
<td>40.5% -</td>
<td>60.75% -</td>
<td>72.9% -</td>
</tr>
<tr>
<td>± 20%</td>
<td>0.8% -</td>
<td>1.6% -</td>
<td>4% -</td>
<td>8% -</td>
<td>16% -</td>
<td>40% -</td>
<td>60% -</td>
<td>72% -</td>
</tr>
<tr>
<td>Accuracy limits</td>
<td>Overall duration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 2%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 4%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 6%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 9%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 11%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 12%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>± 13%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 15.
Upper and Lower Absolute Accuracy Limits of ±1% to ±20% Calculated for Overall Durations Ranging from 1% to 90%. The Grey Section Indicates the Limits Assumed to be Useful in Applied Settings.
| ± 14%  | 0% – 15% | 0% – 16% | 0% – 19% | 0% – 24% | 0% – 34% | 36% – 64% | 61% – 89% | 76% – 100% |
| ± 15%  | 0% – 16% | 0% – 17% | 0% – 20% | 0% – 25% | 5% – 35% | 35% – 65% | 60% – 90% | 75% – 100% |
| ± 16%  | 0% – 17% | 0% – 18% | 0% – 21% | 0% – 26% | 4% – 36% | 34% – 66% | 59% – 91% | 74% – 100% |
| ± 17%  | 0% – 18% | 0% – 19% | 0% – 22% | 0% – 27% | 3% – 37% | 33% – 67% | 58% – 92% | 73% – 100% |
| ± 18%  | 0% – 19% | 0% – 20% | 0% – 23% | 0% – 28% | 2% – 38% | 32% – 68% | 57% – 93% | 72% – 100% |
| ± 19%  | 0% – 20% | 0% – 21% | 0% – 24% | 0% – 29% | 1% – 39% | 31% – 69% | 56% – 94% | 71% – 100% |
| ± 20%  | 0% – 21% | 0% – 22% | 0% – 25% | 0% – 30% | 0% – 40% | 30% – 70% | 55% – 95% | 70% – 100% |
RESULTS

Analyses. To show the representativeness of the \( P_O \) values (the number of MTS samples in which the behaviour was observed divided by the total number of MTS samples), the following data are presented in two ways. The first is as a percentage of samples that fell within accuracy limits, which was determined by calculating absolute and relative accuracy limits for behaviours of varying overall duration (Tables 14 and 15). MATLAB® code that scanned all the output data from the simulated MTS reported the number of samples that fell within each accuracy limit and wrote the number into a Microsoft Excel (2010) spread sheet (Appendix 3). The number of samples that fell within each accuracy limit was turned into a percentage by dividing by the total number of samples and multiplying by 100. To be considered to be representative, 95% of the obtained \( P_O \) values needed to be within the accuracy limits. For low-duration behaviours (i.e., \( \leq 5\% \) overall duration), although the lower limit of the absolute accuracy limits was sometimes 0, obtained \( P_O \) values of 0 were not recorded as being within the accuracy limits. Samples in which behaviour was not detected (i.e., \( P_O = 0 \)) cannot be representative.

Second, some of the data are presented in bubble plots. In the bubble plots, the diameter of each data point (\( \phi_B \)) represents the number of data points (\( N \)) that were the corresponding \( y \)-axis value (obtained percentage of session, \( P_O \)). The number of samples in which each \( P_O \) value was obtained (\( N \)) was found by using the 'remove duplicates' function in Microsoft Excel (2010). Equation 8 (based on the equation used to calculate diameter when the area of the circle is known) was then applied to each of the novel \( P_O \) values and their corresponding frequency of occurrence (\( N \)).

\[
\phi_B = \left( \frac{\sqrt{N}}{\pi} \right) / 10
\]

Equation 8
Bubble plots show the range of obtained percentages ($P_O$) and an estimated probability of each $P_O$ value being obtained (larger diameters of data points show $P_O$ values that were obtained more often).

The data presentation addresses variables (e.g., overall duration, bout duration, MTS interval duration) one at a time, with the values of the other variables selected to best show the effects of the variable of interest. Figures 31, 32, and 33 show criterion records for 60-min examples of simulated low- (1%, 2%, and 5%), intermediate- (10% and 20%), and high-duration behaviours (50%, 75%, and 90%).
Figure 31. Example criterion records for 1-hr extracts of low-duration behaviours (1%, 2%, and 5% overall duration) for varying bout durations. The vertical bars show the seconds during which the behaviour occurred.
Figure 32. Example criterion records for 1-hr extracts of intermediate-duration behaviours (10% and 20% overall duration) for varying bout durations. The vertical bars show the seconds during which the behaviour occurred.
Figure 33. Example criterion records for 1-hr extracts of high-duration behaviours (50%, 75%, and 90% overall duration) for varying bout durations. The vertical bars show the seconds during which the behaviour occurred.
**Bubble plots.** The bubble plots presented in Figures 34 to 37 show the range and the relative number of times each $P_O$ value was obtained with 10-s MTS. A greater likelihood of obtaining representative samples is indicated by a smaller range of $P_O$ values and a large percentage of $P_O$ values that are close to the overall true percentage of time. The overall true percentage of time is indicated by a dashed horizontal line in Figures 34 to 37. Regardless of overall duration, the range of $P_O$ values was greater for longer bout durations than for shorter bout durations, particularly for shorter session durations (e.g., 10-min session duration for behaviours of 50% overall duration, Figure 36). For example, for behaviour of 50% overall duration and a session duration of 10 min, the range of $P_O$ values was 36.7% to 65% for short bout duration, and 0% to 98.3% for very long bout durations. The data also showed that regardless of overall duration, 10-s MTS could produce both under- and over-estimations of overall duration.

For lower-duration behaviours (1% and 10% overall duration), shorter session durations (60 min or less) were likely to produce large underestimations of overall duration, or $P_O = 0$ (i.e., behaviour was not detected), particularly for longer bout durations (Figures 34 and 35). For behaviours of 10% overall duration and very long bout durations (600 s to 1,200 s), samples were likely to underestimate overall duration in session durations shorter than 230 min (3.8 hr), and overestimate overall duration in session durations longer than 230 min. For example, 35% of samples were $P_O = 0$ when session duration was 120 min, and 32% of samples were $P_O = 12.3\%$ when session duration was 390 min. $P_O$ values (i.e., the percentage of samples at each obtained $P_O$ value) were more evenly distributed for higher-duration behaviours (50% and 90% overall duration), indicating that a range of $P_O$ values were equally likely. However, for very long bout durations, behaviours of 50% overall duration were more likely to be underestimated when session duration was 10 min (e.g., 47% of samples were $P_O = 0$) (bottom panel, Figure 36). Behaviours of 90% overall duration were
more likely to be overestimated when session duration was 10 min (e.g., 48% of the samples were $P_O > 90\%$ (bottom panel, Figure 37). Figures 34 to 37 show that shorter bout durations are more likely to be sampled more representatively than longer bout durations.

Figure 34. Effect of bout duration on representativeness. Bubble plot showing samples of 10-s MTS with session durations ranging from 10 min to 480 min, extracted from 100 simulations of behaviour occurring for 1% of the time in short, medium, and long bout durations. Bubble size is relative to the number of samples that produced that value (i.e., larger bubbles indicate a higher percentage of samples).
Figure 35. Effect of bout duration on representativeness. Bubble plot showing samples of 10-s MTS with session durations ranging from 10 min to 480 min, extracted from 100 simulations of behaviour occurring for 10% of the time in short, medium, long, and very long bout durations. Bubble size is relative to the number of samples that produced that value (i.e., larger bubbles indicate a higher percentage of samples).
Figure 36. Effect of bout duration on representativeness. Bubble plot showing samples of 10-s MTS with session durations ranging from 10 min to 480 min, extracted from 100 simulations of behaviour occurring for 50% of the time in short, medium, long, and very long bout durations. Bubble size is relative to the number of samples that produced that value (i.e., larger bubbles indicate a higher percentage of samples).
Figure 37. Effect of bout duration on representativeness. Bubble plot showing samples of 10-s MTS with session durations ranging from 10 min to 480 min, extracted from 100 simulations of behaviour occurring for 90% of the time in medium, long, and very long bout durations. Bubble size is relative to the number of samples that produced that value (i.e., larger bubbles indicate a higher percentage of samples).

**Overall duration.** The effects of overall duration of behaviour on representativeness are shown in Figures 38 and 39. The data presented are from extracted 10-s MTS. Figure 38 shows the percentage of samples that fell within the least stringent absolute accuracy limits (±5%) for low-duration behaviours (1% to 5%) with short bout duration (bout duration of 1 s to
10 s). The data show that representativeness increased with increased overall duration (e.g., in a session 10 min in duration, 44%, 70%, and 97% of the samples were within ± 5% absolute accuracy limits for behaviours of 1%, 2%, and 5% overall duration, respectively). However, as session duration was increased, all low-duration behaviours were likely to be sampled representatively (i.e., at least 91% and 98% of the samples were within ± 5% absolute accuracy limits for all low-duration behaviours when session duration was 40 min and 60 min, respectively).

![Figure 38](image.png)

**Figure 38.** Effect of overall duration on representativeness. The percentage of samples within ± 5% absolute accuracy limits for low-duration behaviours with short bout durations. Samples were extracted using MTS with 10-s intervals in session durations of 10 min to 60 min.

Higher-duration behaviours (i.e., 75% and 90% overall duration) of medium bout duration (30 s to 90 s) were more likely to be sampled representatively with 10-s MTS than were intermediate-duration behaviours (Figure 39). Medium bout duration was chosen for Figure 39 for consistency across intermediate- and high-duration behaviours (i.e., short bout duration was not possible for 90% overall duration). At a minimum session duration of 10
min, 86% and 99% of samples were within the least stringent ± 10% absolute accuracy limits for behaviours of 75% and 90% overall duration. However, 51%, 59%, and 61% of samples were within ± 10% absolute accuracy limits for behaviours of 10%, 20%, and 50% overall duration, respectively. Session duration of at least 40 min was required for more than 90% of the samples to fall within the ± 10% absolute accuracy limits for all intermediate- and high-duration behaviours. Higher-duration behaviours were more likely to be sampled representatively than low-duration behaviours.

![Figure 39](image)

*Figure 39.* Effect of overall duration on representativeness. The percentage of samples within ± 10% absolute accuracy limits for intermediate- and high-duration behaviours with medium bout durations. Samples were extracted using MTS with 10-s intervals in session durations of 10 min to 60 min.

**Bout duration.** The effects of bout duration on representativeness are illustrated in Figures 40 and 41. The data presented are from extracted 10-s MTS. Figure 40 shows the percentage of samples of low-duration behaviours (1% to 5% overall duration) that fell
within the least stringent absolute accuracy limits (± 5%) for short (1 s to 10 s), medium (30 s to 90 s), and long (120 s to 300 s) bout durations.

Figure 40. Effect of bout duration on representativeness. The percentage of samples within ±5% absolute accuracy limits for low-duration behaviours (1% to 5%). Samples were extracted using MTS with 10-s intervals in session durations of 10 min to 60 min.

Figure 41 shows the percentage of samples of intermediate- (10% and 20% overall duration) and high-duration behaviours (50% to 90% overall duration) that fell within the least stringent absolute accuracy limits (± 10%) for short (1 s to 10 s), medium (30 s to 90 s), long (120 s to 300 s), and very long (600 s to 1,200 s) bout durations. The data show that representativeness decreased with increases in bout duration, across all overall durations (1% to 90%) (e.g., in a 10-min session, 95%, 71%, 43%, and 26% of the samples were within ±10% absolute accuracy limits for intermediate-and high-duration behaviours of short, medium, long, and very long bout duration, respectively).
Figure 41. Effect of bout duration on representativeness. The percentage of samples within ±10% absolute accuracy limits for intermediate- and high-duration behaviours (10% to 90%). Samples were extracted using MTS with 10-s intervals in session durations of 10 min to 60 min.

In addition, low-duration behaviours with medium or long bout durations were less likely to be sampled representatively than higher-durations behaviours (≥10% overall duration) with medium or long bout duration. For example, for low-duration behaviours with medium bout duration, the percentage of samples within ±5% absolute accuracy limits increased from 11% to 67% when session duration was increased from 10 min to 60 min (Figure 40). However, for higher-duration behaviours with medium bout duration, the percentage of samples within the least stringent ±10% absolute accuracy limits increased from 71% to 99% when session duration was increased from 10 min to 60 min. Shorter bout durations were more likely to be sampled representatively than longer bout durations.

Low-duration behaviours. When the requirements that 60 min was the maximum session duration and ≥95% of the samples were within ±5% absolute accuracy limits were
implemented, low-duration behaviours were not sampled representatively. Figure 42 indicates the 60 min maximum session duration (vertical dashed line) and the point at which 95% of the samples are within ±5% absolute accuracy limits (horizontal dashed line). Medium (30 s to 90 s) and long (120 s to 300 s) bout durations were never sampled representatively within the maximum acceptable session duration (i.e., less than 95% of the samples were within ± 5% absolute accuracy limits) (middle and right panels, Figure 42). For behaviours of 1% overall duration, only 10-s MTS in 60-min sessions produced representatives samples (top left panel, Figure 40). Behaviours of 2% and 5% overall duration could be sampled representatively sessions of ≤ 60-min, but only by MTS ≤ 20-s intervals. Figure 43 shows that intermediate- and high-duration behaviours were more likely to be sampled representatively within 60-min sessions, but not for all bout durations and with all MTS interval durations. Figures 42 and 43 show that low-duration behaviours (occurring for overall ≤ 5%) are only likely to be sampled representatively within a very narrow set of values of the dimensions of behaviour and parameters of the measurement system (e.g., overall duration, bout duration, session duration, and MTS interval duration). Henceforth, low-duration behaviours (≤ 5%) are removed from the data analysis.
Figure 42. Percentage of samples within ±5% absolute accuracy limits extracted from simulations of behaviours of 1% (top panels), 2% (middle panels), and 5% (bottom panels) overall duration, and short, medium, and long bout durations. Samples were extracted using MTS with intervals ranging from 10 s to 3,600 s (60 min). The horizontal dashed line shows the point at which 95% of the samples were within the limits, and the vertical dashed line shows session duration of 60 min.
Figure 43. Percentage of samples within ±10% absolute accuracy limits extracted from simulations of behaviours of 10% (top panels), 50% (middle panels), and 90% (bottom panels) overall duration, and short, medium, long, and very long bout durations. Samples were extracted using MTS with intervals ranging from 10 s to 3,600 s (60 min). The horizontal dashed line shows the point at which 95% of the samples were within the limits, and the vertical dashed line shows session duration of 60 min.
**MTS interval duration.** The mean percentage of samples within ± 10% absolute accuracy limits was calculated by dividing the sum of the percentage of samples within ± 10% absolute accuracy limits by the number of bout durations analysed (four for 10% to 75% overall duration, three for 90% overall duration). Figure 44 shows that MTS with intervals of 300 s or longer never produced samples representative of behaviours of 10% to 90% overall duration. The longer the MTS interval, the smaller the mean percentage of samples within ± 10% absolute accuracy limits. For example, at the maximum acceptable session duration of 60 min for behaviours of 10% overall duration, 48.2%, 28.5%, 0%, and 0% of MTS samples were within ± 10% absolute accuracy limits when interval duration was 300-s, 600-s, 1,800-s, and 3,600-s intervals respectively. A larger mean percentage of samples fell within ± 10% absolute accuracy limits for higher-duration behaviours (e.g., 90.3% of 300-s MTS samples in a 60-min session for behaviours of 90% overall duration). However, averaged across all bout durations, the mean percentage did not meet the requirement that ≥ 95% of the samples fell within the limits. Henceforth, MTS with intervals of 300 s or longer are removed from the data analysis.

**Longer bout durations.** Long (120 s to 300 s) and very long (600 s to 1,200 s) bout durations were not sampled representatively for intermediate- and high-duration behaviours (Figure 45). For example, when 10-s MTS was used to sample behaviours with long bout durations in 60- min sessions, 77%, 82%, 78%, and 95% of the samples were within ± 10% absolute accuracy limits for behaviours of 10%, 20%, 50%, and 75% overall duration, respectively. Although 95% of the samples were within ± 10% absolute accuracy limits for behaviours of 75% overall duration when session duration was 60 min, session durations of ≤ 50 min produced fewer than the acceptable percentage of samples within the limits (95%). For behaviours of 90% overall duration, ≥ 95% of the samples were within ± 10% absolute
accuracy limits when session duration was 20 min or longer, regardless of MTS interval duration (10 s to 60 s).

Figure 44. Mean percentage of samples within ±10% absolute accuracy limits extracted from simulations of behaviours of overall durations ranging from 10% to 90% across all bout durations. Samples were extracted using MTS with intervals ranging from 300 s to 3,600 s (5 min to 60 min). The horizontal dashed line shows the point at which 95% of the samples were within the limits, and the vertical dashed line shows session duration of 60 min.
Figure 45. Percentage of samples within ±10% absolute accuracy limits extracted from simulations of behaviours of overall durations ranging from 10% to 90% for long and very long bout durations (120 s to 300 s, and 600 s to 1,200 s, respectively). Samples were extracted using MTS with intervals ranging from 10 s to 60 s. The horizontal dashed line shows the point at which 95% of the samples were within the limits, and the vertical dashed line shows session duration of 60 min.

When bout duration was very long, 10%, 36%, 42%, and 71% of the samples were within ±10% absolute accuracy limits for behaviours of 10%, 20%, 50%, and 75% overall duration.
duration, respectively. Although 97% of the samples were within ± 10% absolute accuracy limits for behaviours of 90% overall duration when session duration was 60 min, session durations of 50 min or less produced fewer than the acceptable percentage of samples within the limits. Henceforth, long bout durations for behaviours of 10% to 75% overall duration, and very long bout durations for behaviours of 10% to 90% are removed from the data analysis.

**Three-dimensional (3-D) mesh plots.** The elimination of values of the dimensions of behaviours and sampling methods for which MTS did not produce representative samples resulted in remaining overall durations of 10%, 20%, 50%, 75% with short and medium bout durations, and overall duration of 90% with medium and long bout durations. A 3-D mesh plot was created for each remaining combination of overall duration and bout duration (Figures 46 to 50). For each plot, the session duration at which ≥ 95% of the 100 samples were within accuracy limits was plotted against the corresponding accuracy limit and MTS interval duration (10 s to 60 s). For example, for 10% overall duration and short bout duration, when MTS with 60 s intervals was used (z-axis), session duration of 240 min (y-axis) is required for ≥ 95% of the samples to fall within ± 4% absolute accuracy limits (x-axis) (right panel, Figure 46). The colours of the mesh plot indicate approximate session durations (e.g., dark blue for shorter sessions of up to 120 min, yellow for sessions between 300 min and 400 min in duration, and red for very long session durations of 480 min). The colours are an approximate guide only; specific session duration values can be determined from the z-axis scale. The practical utility of the 3-D dimensional plots is to identify the required session duration when acceptable accuracy and practical MTS interval duration are selected. In order to read values more accurately from the plots, the graphs sometimes require rotation. Examples of rotations are provided in Appendix 4.
Figure 46. Three-dimensional mesh plot from which required session duration can be determined when acceptable accuracy limits and practical MTS interval duration is selected for behaviours of 10% overall duration. Behaviours with short bout duration (1 s to 10 s) are represented in the left plot, and behaviours with medium bout duration (30 s to 90 s) in the right plot.
Figure 47. Three-dimensional mesh plot from which required session duration can be determined when acceptable accuracy limits and practical MTS interval duration is selected for behaviours of 20% overall duration. Behaviours with short bout duration (1 s to 10 s) are represented in the left plot, and behaviours with medium bout duration (30 s to 90 s) in the right plot.
Figure 48. Three-dimensional mesh plot from which required session duration can be determined when acceptable accuracy limits and practical MTS interval duration is selected for behaviours of 50% overall duration. Behaviours with short bout duration (1 s to 10 s) are represented in the left plot, and behaviours with medium bout duration (30 s to 90 s) in the right plot.
Figure 49. Three-dimensional mesh plot from which required session duration can be determined when acceptable accuracy limits and practical MTS interval duration is selected for behaviours of 75% overall duration. Behaviours with short bout duration (1 s to 10 s) are represented in the left plot, and behaviours with medium bout duration (30 s to 90 s) in the right plot.
Figure 50. Three-dimensional mesh plot from which required session duration can be determined when acceptable accuracy limits and practical MTS interval duration is selected for behaviours of 90% overall duration. Behaviours with medium bout duration (30 s to 90 s) are represented in the left plot, and behaviours with long bout duration (120 s to 300 s) in the right plot.
The 3-D mesh plots show that, in general, longer session durations are required for longer bout durations. For example, the minimum session duration, or the lowest blue part of the mesh plot, is higher up the \( z \)-axis (session duration) on the right plot (medium bout duration) than the left plot (short bout duration) for behaviours of 10% overall duration (Figure 46). When short MTS interval durations (e.g., 10 s) were used for some behaviours across sessions of less than the maximum 480 min, \( \geq 95\% \) of the samples fell within the most stringent accuracy limits (\( \pm 1\% \)). Therefore, there are some gaps in the mesh plot for some of the overall durations (e.g., 90% overall duration and long bout duration, right panel, Figure 46).

The mesh plots were created in SigmaPlot 11.0 (Systat Software Inc.). For each overall duration, at some MTS interval durations, there were absolute accuracy limits within which \( \geq 95\% \) of the samples did not fall, regardless of increases in session duration to the maximum 480 min. Although the graphing software extrapolated the data (i.e., plotted data showing that longer sessions could produce data within the limits), this resulted in the 3-D mesh plots generalising the data to values that were not obtained. To illustrate the difference in the plots when the extrapolated data were removed, Figure 51 shows the accuracy limits within which \( \geq 95\% \) of the samples did not fall (compare with Figure 48). For example, when MTS with 60-s intervals was used to measure behaviour of 50% overall duration and short bout duration, \( \geq 95\% \) of the samples were within absolute accuracy limits of \( \pm 5\% \) and above. Session duration of 290 min was required for samples to fall within \( \pm 5\% \) absolute accuracy limits (Figure 49). Increases in session duration never resulted in \( \geq 95\% \) of the samples falling within absolute accuracy limits of \( \leq \pm 4\% \). However, to show these data (i.e., create the peaks in the mesh plot), required sessions durations of zero were entered for absolute accuracy limits within which \( \geq 95\% \) of the samples never fell. The result is potentially misleading data (indicated by an arrow on the right panel of Figure 51).
Therefore, although Figures 46 to 50 contain extrapolated data, they are less likely to be misleading.

**Three-dimensional (3-D) stem plots.** Should more accurate values be required from the 3-D plots, 3-D stem plots may provide a more detailed alternative. The same variables were plotted in as in the mesh plots, but the data remained as individual values rather than being combined into a smooth plot. Figure 52 provides an example of the 3-D stem plot for behaviour of 50% overall duration. Specific values can be more easily determined, e.g., 60-s MTS across a 120-min session will produce data likely to be within ±7% accuracy limits for a behaviour occurring for 50% overall with medium bout durations (right panel, Figure 52). In addition, the 3-D stem plots show values for which data are not likely to fall within the specified absolute accuracy limit. For example, MTS with 60-s intervals will not produce data within absolute accuracy limits of ≤±4% for behaviours of 50% overall duration and short bout duration (left panel, Figure 52). Stem plots for behaviours of 10%, 20%, 75%, and 90% overall duration can be found in Appendix 4.

The 3-D mesh and stem plots represent a data-based method for selecting sampling method, and provide 'decision rules' for obtaining representative samples. For example, estimations of overall and bout duration, choosing acceptable error and a practical MTS interval duration, can be used to select the required observation session duration.
Figure 51. Three-dimensional mesh plot from which required session duration can be determined when acceptable accuracy limits and practical MTS interval duration is selected for behaviours of 50% overall duration. Extrapolated data are eliminated to show the session durations and MTS interval durations for which 95% of more of the samples did not fall within the accuracy limits. Behaviours with short bout duration (1 s to 10 s) are represented in the left plot, and behaviours with medium bout duration (30 s to 90 s) in the right plot.
Figure 52. Three-dimensional stem plot from which required session duration can be determined when acceptable accuracy limits and practical MTS interval duration is selected for behaviours of 50% overall duration. Behaviours with short bout duration (1 s to 10 s) are represented in the left plot and behaviours with medium bout duration (30 s to 90 s) in the right plot.
ANALYSIS OF DECISION RULES

Selection of Study 1 behaviours. Full-day data from Study 1 were used to test the data-based method for selecting sampling method (i.e., MTS interval duration and session duration) illustrated in Figures 46 to 50. Daily true values that were within 1% of a value of overall duration analysed in Study 2 were identified. There were no behaviours of 75% or 90% overall duration, but behaviours of 10%, 20%, and 50% overall duration were identified (Table 16). For the four selected behaviours, mean and median bout durations were determined using the variable summary statistics function, and bout range was determined using the bout distribution function in ObsWin32. Figure 53 shows the range and distribution of bout durations for the four selected behaviours. Each behaviour selected from Study 1 was sampled using the decision rules that most closely matched the median and mode bout duration. For example, attention for Participant 3 on Monday occurred for 10.8% of the day, with median bout duration of 12 s and mode bout duration of 4 s (Table 16). Therefore, samples were extracted using the decision rule data for behaviours occurring with 10% overall duration and short bout duration. Table 16 shows which decision rules were used to sample each of the behaviours.
<table>
<thead>
<tr>
<th>Participant</th>
<th>Day</th>
<th>Behaviour</th>
<th>Overall duration</th>
<th>Bout duration (s)</th>
<th>Number of occurrences</th>
<th>Rules with which tested</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>Median</td>
<td>Range</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Toe-walking</td>
<td>50.3%</td>
<td>110.2</td>
<td>65</td>
<td>4 - 870</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>On-task</td>
<td>20.3%</td>
<td>138.8</td>
<td>58</td>
<td>3 – 623</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>Attention</td>
<td>19.6%</td>
<td>32.3</td>
<td>12</td>
<td>2 – 449</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Attention</td>
<td>10.8%</td>
<td>25.4</td>
<td>12</td>
<td>2 - 176</td>
</tr>
</tbody>
</table>
Sampling. For each behaviour, the range of session durations, MTS interval durations, and absolute accuracy limits dictated by the corresponding decision rule was tested. For each combination of selected MTS interval duration and absolute accuracy limits, the required session duration was extracted using the time filter and MTS sampling conversion functions in ObsWin32. The resulting obtained percentage of time in which the behaviour was recorded as occurring ($P_o$) was compared to the true daily percentage (Figure 54).

*Figure 53.* Range and distribution of bout durations of Study 1 behaviours selected for decision rule testing. Note the change in $x$-axis scale across panels.
Figure 54 shows the true daily percentage (solid horizontal lines) and the least stringent absolute accuracy limits (± 10%) for each tested behaviour. A distinction was made between obtained percentage values that fell within the accuracy limits within which they were expected to fall based on the decision rule (filled circles), and those that did not (open circles). $P_O$ values could fall within the least stringent ± 10% absolute accuracy limits but outside the prescribed accuracy limit within which they were predicted to fall.

For the behaviour of 50.3% overall duration and medium bout duration, 35% of the $P_O$ values fell within the prescribed accuracy limits, and 47.5% of the $P_O$ values fell within the least stringent ± 10% accuracy limits. More $P_O$ values fell within the prescribed accuracy limits for the lower-duration behaviours (47.5%, 63.4%, and 72.1% for behaviours of 20.3% overall duration with medium bout durations, 19.6% overall duration with short bout durations, and 10.8% overall duration with short bout durations, respectively). Behaviours of 19.6% and 10.8% overall duration with short bout durations were most often sampled representatively, with 90.7% and 80.5% of the $P_O$ values falling within the ± 10% accuracy limits, respectively. Across all four behaviours, behaviour was not detected ($P_O = 0$) in only 2.5% of the extracted samples.

Although the behaviours from Study 1 were matched to the decision rules based on overall duration and bout duration, Figure 55 shows the differences in the distribution of the behaviours and variability in bout duration. Higher-duration behaviours (50.3% and 20.3% overall duration) with medium bout duration differed most from the simulated data in that bout durations and interbout intervals (i.e., the duration of time between occurrences of behaviour) were more variable than the simulated data (top two panels, Figure 55). Table 16 also shows that the behaviours with longer-bout durations had larger ranges of bout duration. Behaviours with more variable bout durations and interbout intervals were less likely to be sampled representatively using the decision rules developed in Study 2.
Figure 54. Obtained percentages of time of behaviours with which Study 2 decision rules were tested. Filled and open circles indicate obtained percentages that did and did not fall within the absolute accuracy limits prescribed by the decision rules, respectively. The solid lines represent the true percentage of the behaviour, and the dashed lines indicate the ±10% absolute accuracy limits. Note the change in y-axis scale across panels.
Figure 55. Criterion records for 1-hr extracts of Study 1 data for the behaviours with which the decision rules were tested compared with 1-hr extracts of simulated Study 2 data. The vertical bars show the seconds during which behaviour occurred.
DISCUSSION

Each combination of behaviours of low (1%, 2%, and 5%), intermediate (10% and 20%), and high overall duration (50%, 75%, and 90%) with varied bout duration (1 s to 10 s, 30 s to 90 s, 120 s to 300 s, and 600 s to 1,200 s) was simulated 100 times. From the simulated 8-hr streams of behaviour, MTS samples with intervals of 10 s, 15 s, 20 s, 30 s, 60 s, 300 s, 600 s, 1,800 s, and 3,600 s were extracted across a range of session durations from 10 min to 480 min (8 hr).

Study 2 produced findings regarding values of dimensions of behaviour and parameters of the sampling method for which MTS does not produce representative samples. The main findings were that MTS did not produce representative samples when: 1. behaviours were ≤ 5% overall duration; 2. bout durations were greater than 120 s; and, 3. MTS interval duration was ≥ 300 s. More representative samples were obtained for: 1. higher duration behaviours (e.g., behaviours of ≥ 10% overall duration); and, 2. short and medium bout durations (e.g., behaviours occurring for between 1 s and 90 s per occurrence). The higher the overall duration of behaviour and the shorter the bout durations, the more likely a sample was to be representative of true overall duration.

From the data obtained from simulating behaviours and different parameters of MTS (i.e., interval duration and session duration), data-based decision rules for behaviours of 10%, 20%, 50%, 75% and 90% overall duration with short or medium bout durations (and long bout duration for 90% overall duration only) were developed. The decision rules dictated required session duration when overall and bout duration were estimated, and acceptable accuracy and practical MTS interval duration were identified. To test whether the use of the decision rules resulted in representative samples of behaviour, the decision rules were used to select sampling parameters for behaviours measured in Study 1. The results showed that the
decision rules were more likely to produce representative samples of behaviours with short and less variable bout durations.

**DIMENSIONS OF BEHAVIOUR**

**Overall duration.** Whereas previous studies have assessed the representativeness of samples of durations of behaviour selected arbitrarily, the current study selected a range of durations based on the range and distributions reported in a sample of the behavioural literature. The results of the Study 2 corroborate previous findings that higher-duration behaviours are more likely to be sampled representatively than lower-durations behaviours (e.g., Powell, Martindale, Kulp, Martindale, & Bauman, 1977). However, there are few studies that have evaluated the effect on representativeness of differing behavioural duration, and those that have only tested a small range of intermediate- and higher-duration behaviours. Powell et al. evaluated behaviours of 20%, 50%, and 80% overall duration. However, the analysis of reported durations in the *Journal of Applied Behaviour Analysis* showed that duration measures have been used for a larger range of durations, including lower-duration behaviours (i.e., behaviours occurring for < 10% of the session).

Gardenier, MacDonald, and Green (2004) evaluated the representativeness of MTS for stereotypy of approximately 5% to 55% overall duration. Their categorisations of low-, intermediate-, and high-duration behaviours were similar to those in the current study (i.e., overall duration of ≤ 19% were low-duration, 20% to 39% intermediate-duration, and ≥ 40% high-duration). Gardenier et al. found that although MTS with 10-s intervals produced reasonably representative samples of behaviour regardless of overall duration, error (i.e., the difference between true and obtained values) was larger for low-duration behaviours. However, within the category of low-duration stereotypy, no analysis was conducted on the representativeness of very low durations (i.e., ≤ 5%, as conducted in the current study),
versus the representativeness of higher-duration behaviours (i.e., 10% to 19%). The current study found that low-duration behaviours were not sampled representatively by MTS, even with short MTS intervals.

The inability of MTS to produce representative samples was reported by Saudargas and Zanolli (1990), who demonstrated that low-duration behaviours with short bout durations were not detected at all in some samples obtained with 15-s MTS. However, their data compared duration measures obtained with continuous computer recording and 15-s MTS, but did not compare either to true values (i.e., a measure of representativeness). It is also important to note that most studies of the representativeness of measures of behaviour duration have used short sessions that have been videotaped (e.g., Gardenier et al., 2004; 10 min sessions) or in which the behaviour has been contrived (e.g., Powell et al., 1977; 30-min sessions). The current study evaluated a range of session durations and behaviours for which overall duration was distributed across a simulated 8-hr day. Both of these factors are discussed in more detail below.

**Low-duration behaviours.** The results of Study 2 support the suggestion by Alevizos, DeRisi, Liberman, Eckman, and Callahan (1978) that low- and high-duration behaviours are unlikely to be sampled representatively with the same measurement system. In Study 2, low-duration behaviours (≤ 5% overall duration) were unlikely to be sampled representatively across ≤ 60-min observation sessions, particularly with MTS intervals longer than 10 s. The finding that low-duration behaviours (≤ 5% overall duration) are not sampled representatively with MTS leads to consideration of how they may be sampled representatively.

**Practical considerations for measuring low-duration behaviours.** The solution to measuring low-duration behaviours representatively is likely to be affected by the purpose of data collection and the dimension of interest. Potential methods of representatively
measuring low-duration behaviours may include: 1. MTS with intervals shorter than 10 s; 2. frequency recording; 3. continuous recording; 4. collecting data to identify times of the day in which to conduct observations, and; 5. automated measurement.

MTS with very small intervals (i.e., 5 s) has been suggested for low-duration behaviours (Saudargas & Zanolli, 1990), but there have been few empirical analyses of very short MTS intervals which, anyway, are likely to be impractical in many settings. Frequency recording may be a satisfactory recording method if the behaviour occurs in discrete events (i.e., essentially zero duration), or if bout duration is short (1 s to 10 s) and invariable, or even long and invariable. Alternatively, continuous recording has been suggested to be useful in measuring low-duration behaviours with short bout duration (Daigle & Siegford, 2013). However, continuous recording over long sessions is unlikely to be practical, and perhaps is best used when the time at which the behaviour occurs can be easily identified. Observation sessions can then be conducted at cost-effective, specific times (although this will not necessarily provide a representative measure of overall duration).

For low-duration behaviours, the distribution of occurrences across the day may be of particular importance. For example, for behaviour of low overall duration that occurs in bursts of multiple bouts with long interburst intervals between, a scatter plot to determine the approximate distribution of behaviour across a day could be used to identify times of the day to be sampled. A scatter plot divides the day into large blocks, within which behaviour is recorded as occurring or not occurring. The resulting data indicate times of the day during which the behaviour is likely to occur (i.e., an indication of stimulus control), and any patterns within or across days (e.g., behaviour is uniformly distributed across days or weeks). One potential advantage of using a scatter plot is the ease with which they can be completed by staff or carers in applied settings due to the low response effort required (Touchette, MacDonald, & Langer, 1985). The time-of-interest may then be reduced to the times of day
during which the behaviour is likely to occur rather than a full day or week, which increases
the overall duration (relative to the time-of-interest), and allows a sampling method to be
selected based on the revised estimate of duration. Similarly, a scatter plot allows a
practitioner to identify potential antecedents or consequences of a behaviour (McGill, 1999),
providing further assessment information to be gathered and used to inform intervention. The
purposes of data collection and the dimension of interest must also be considered, because it
may be that the stimulus conditions under which behaviour occurs or the bout durations are
of more relevance to effecting behaviour change than the overall duration of the behaviour.

Another alternative for measuring low-duration behaviours may be automated
measurement. Crowley-Koch and Van Houten (2013) discussed the development of
technology such as accelerometers, the data from which can be sent to and analysed using
smart phone applications or websites, and how such automated methods of measuring
behaviours could be adapted for behaviour analysis. For example, Yoder, Oller, Richards,
Gray, and Gilkerson (2013) recorded the vocalisations of children with autism using portable,
wearable digital language processors. Calibrated correctly and with appropriately set
thresholds for recording the occurrence of behaviour (e.g., minimum force or volume),
automated measurement systems may produce representative samples of low-duration
behaviours. Automated measurement systems are capable of recording data for long times-
of-interest (e.g., whole days; Yoder et al.), and remove the requirement of a dedicated
observer. Goodwin, Velicer, and Intille (2008) acknowledged that despite other advantages
of automated measurement such as removing observer and discontinuous sampling errors,
and the ability to produce large amounts of behavioural and environmental data, exploratory
analyses of automated measurement devices are required. Such analyses would identify
potential user errors, thresholds below which behaviour is not recorded, and issues in
calibrating the devices.
**Bout duration.** Few studies have evaluated the effects of bout duration on representativeness of samples, or the interaction between bout duration and overall duration. Powell (1984a) kept the overall duration of simulated behaviour constant at 50% (of 30-min sessions), whilst varying the duration of each occurrence (bout duration). The ratio of lower to upper bout limits was constant at 1:3, and a range of 1 s and 3 s to 81 s and 243 s of bout durations was simulated. Increases in bout duration resulted in decreases in the representativeness of partial-interval recording samples. The results of Study 2 corroborated these findings, and showed that representativeness continues to decrease with increases in bout duration beyond those analysed by Powell.

Wirth, Slaven, and Taylor (2014) simulated bout durations ranging from 1 s to 256 s, but they kept bout duration constant within each simulation (e.g., there was no range, all bouts were of equal duration). They reported that when the MTS interval was equal to or less than the bout duration, representative samples were obtained. Upon inspection of the figures, their data showed increases in representativeness when event duration increased in a 60-min observation session. The difference in effect of bout duration between the study by Wirth et al. and Study 2 may be accounted for by several factors. First, Wirth et al. kept the bout duration of each event constant in each simulation, whereas bout duration was varied in Study 2. Second, Wirth et al. analysed maximum bout duration of 256 s and a minimum overall duration of 8%, whereas Study 2 evaluated longer bout durations and lower overall durations that produced poorly representative samples. Third, Wirth et al. did not discuss the interaction between bout duration and overall duration, rather focusing on the interaction between bout duration and MTS interval duration. Therefore, their conclusion that representative samples can be obtained when the MTS interval duration is shorter than the bout duration is too general. The data from Study 2 show that short MTS interval durations are unlikely to produce representative samples when the overall duration of behaviour is low,
or when bout duration is long (i.e., ≥ 120 s). Perhaps the differences in conclusions from Wirth et al. and the current study show that pairwise comparisons of variables should be complemented with analyses of the effects of all the variables of interest, because the relationship between two variables is unlikely to be unaffected by other variables. In addition, Wirth et al. acknowledged that their simulations confounded bout duration with overall duration as the variables were not evaluated separately. In addition, their events were able to overlap (i.e., an event could begin in the second consecutive to the last second of the previous event), whereas the events simulated in Study 2 were separated by at least 1 s. Overlapping events could result in longer bout durations and alter the distribution of events across the session.

Some studies have suggested that MTS intervals shorter than the lower limit of the bout durations will produce representative samples, and therefore an estimate of bout duration can guide the sampling method selected (Ary & Suen, 1983). The task analysis of practical steps to obtain representative samples of behaviour using MTS provided by Ary and Suen is a useful tool for practitioners. However, although Study 2 demonstrated that MTS intervals shorter than the lower limit of the bout duration could produce representative samples, the data also showed that the variability in bout durations affected the representativeness of MTS.

**Distribution of behaviour.** Tiger et al. (2013) suggested that consideration of the variability of behaviour within a day (i.e., distribution) may be of greater concern than overall duration when the representativeness of samples is considered. Tiger et al. characterised behaviours as high- or low-variability based on the standard deviation of each of the 10-min continuous observation samples they extracted from criterion records. The high-variability behaviours measured by Tiger et al. were likely to be distributed unevenly across the day. Distribution is related to the locus in time of each occurrence (i.e., affected by the interbout
intervals, or the time between occurrences). Highly variable interbout intervals are likely to reduce the representativeness of samples of behaviour. An infinite number of distributions of behaviour across a day are possible, and the results of Study 2 showed that any distribution including variable interbout intervals results in less representative sampling.

The finding that the decision rules generated in Study 2 did not produce representative samples for all behaviours selected to be sampled from Study 1 is likely related to the differences in the distribution of behaviour. Bout durations and interbout intervals were randomly selected by the simulations in Study 2, and because bout durations were more variable in the data collected in applied settings in Study 1, the interbout intervals were also more variable. This finding corroborates the findings of Tiger et al. (2013). Similarly, Paquet, Punnett, Woskie, and Buchholz (2005) found that more samples of construction workers' behaviours (e.g., lateral bending, lifting, hand-tool use) were required when the bout duration of behaviours was variable within a day. In studying the durations of the behaviours of chimpanzees, Thiemann and Kraemer (1984) also found that behaviours that varied within a day required more samples to obtain representative data regarding duration. Particular behaviours more likely to be variable were identified by their function and environmental factors (e.g., aggression could not occur in the absence of other chimpanzees). Research on the distribution of behaviours that are more likely to be variable (e.g., perhaps on-task behaviour is more variable across a day than adult attention) may help practitioners to choose measurement methods based on the effect of distribution on representativeness.

It is important to acknowledge, however, that variability in behaviour (i.e., bout durations and interresponse intervals) is problematic when the overall duration of behaviour is of interest, resulting in more required samples (observation sessions) to obtain a mean that is representative of the true overall duration. Measuring variability in behaviour across or within days may be useful when the purpose of data collection is to determine the antecedent
conditions under which behaviour occurs with higher duration. In such instances, the concern for an observer is that the variability across observation sessions reflects the variability in behavioural duration and is not due to measurement error alone.

A method for determining how much of the variability obtained across samples is due to the variability in behaviour, time of day, and setting is to conduct a Generalizability (G) study, based on generalizability theory. G theory assumes that error is a product of the aspects of a measurement system (e.g., setting, observers) (Hintze, Owen, Shapiro, & Daly, 2000). The purpose is therefore to determine the relative contributions of measurement factors (source of variability) to error in the measurement. In a G study, variance components for each source of variability are determined through repeated measures ANOVA (Hintze et al., 2000). Following the estimation of variance components, a practitioner can choose to report the variance components, attempt to reduce the error from a particular source (e.g., observer error), or conduct a Decision (D) study (Suen & Ary, 1989). Whereas a G study analyses current sources of error, a D study allows a practitioner to calculate expected sources of error when sources of variability are manipulated (McWilliam & Ware, 1994).

Hintze and Matthews (2004) conducted a G study to determine the variability across samples of on-task behaviour that could be explained by differences in the setting, time or day, and individual student. The remaining unexplained error may indicate measurement error. Although G studies, and subsequent D studies that facilitate decisions regarding parameters such as the number and duration of observations, have been demonstrated to be useful in directly observing behaviours in applied settings (e.g., in classrooms; McWilliam & Ware, 1994), many practitioners would find the method complicated and impractical. The ability to conduct a repeated-measures analysis of variance (as required for G studies; Volpe, McConaughy, & Hintze, 2009) may be time-consuming, many practitioners would not be
familiar with the method, and the process and results may be difficult to explain to teachers,
carers, or other professionals. Perhaps G and D studies are useful for informing the literature
on behavioural measurement, and further research could evaluate simplifying the process for
use by practitioners.

There have been a few studies that have measured the distribution of behaviour across
a day or longer, or evaluated the effect of distribution on the representativeness of samples.
For example, Landesman-Dwyer, Stein, and Sackett (1978) recorded categories of behaviour
across full days (from early morning to late at night), collecting more than 16,000 hrs of data
across 406 staff and residents in group homes for people with intellectual disabilities. From
their data, they were able to analyse relationships between the overall durations of behaviours
and other variables such as severity of intellectual disability and the size of the home.
Landesman-Dwyer and Sackett were able to use data collected in 24-hr observation sessions
to describe the patterns of sleep and levels of activity in children with profound intellectual
disabilities. Burgio et al. (1994) used continuous recording to measure the duration of the
disruptive vocalisations of people in nursing homes, finding a skewed distribution of
behaviour across the day (i.e., increases in duration across the day). Their data could be
described as diurnally consistent; following a similar pattern each day.

For diurnally consistent behaviour, conducting a short (i.e., 60 min) observation
session at any time during the day is unlikely to produce a sample representative of daily
overall duration. However, if the dimension of interest is to determine the changes in
duration across a whole day, then MTS with longer intervals (i.e., 300 s or 5 min and above)
may produce representative samples. Regardless of bout duration, the data from Study 2
showed that for behaviours of 50% overall duration, MTS with 300 s intervals could produce
representative samples when session duration was approximately 6 hrs or longer. However,
MTS with intervals of 1,800 s (30 min) or 3,600 s (60 min) never produced representative
samples, regardless of session duration or overall duration. In applied settings, conducting one or two whole days of MTS observation to determine the changes in duration of a uniformly distributed behaviour across the day may yield more useful data than multiple hour-long observations at randomly selected times of the day.

An overall conclusion regarding the effect of distribution of behaviour on the representativeness of samples is that obtaining information regarding the likely variation in duration across a day (i.e., through informant methods or scatter plots) may affect the choice of sampling method. If measuring the variation in duration across a day is important, then MTS with longer intervals may be implemented across a small number of whole days. Informant methods or scatter plots could be used to identify times of the day during which the behaviour occurs, and MTS with shorter intervals can be conducted at that time of day to yield a baseline measure of the duration of the behaviour, and subsequently evaluate the effectiveness of an intervention. The purpose of data collection and whether the distribution of behaviour is important will largely affect the choice of sampling method. Categorising patterns of behavioural distribution and measuring their effects on the representativeness of behaviour would benefit from further study.

**SAMPLING METHOD PARAMETERS**

*Momentary time sampling (MTS).* Gardenier et al. (2004) suggested that the circumstances under which errors in MTS could be reduced should be systematically evaluated. Study 2 evaluated systematically the effects of sampling parameters (e.g., MTS interval duration and session duration) and dimensions of behaviour (e.g., overall duration and bout duration) on the representativeness of samples obtained with MTS. The data from Study 2 corroborate the results of studies such as Powell et al. (1977) which showed that the error produced by MTS is unsystematic (i.e., MTS both over- and underestimated duration).
However, MTS can appear to produce representative samples and unsystematic error when the means of errors or obtained values are reported. The advantage of bubble plots to display data is to show the range of possible obtained percentages of duration, although they are more difficult to interpret when the range of obtained values is small and the bubbles overlap. Practitioners are likely to be more concerned about the representativeness of each data point, rather than the mean, because a mean provides little information about the range and whether the variability across data points is due to unrepresentative samples or variability in behaviour.

The bubble plots also indicated that the error produced by MTS is systematic. As found by Wirth et al. (2014), MTS was more likely to underestimate low-duration behaviours (e.g., 10% overall duration) and overestimate high-duration behaviours (e.g., 90% overall duration). However, Study 2 also found that this bias was affected by MTS interval duration (shorter intervals resulted in fewer under- or overestimations) and, for some combinations of overall duration and bout duration, session duration. For example, for behaviours of 10% overall duration and very long bout durations, observation sessions of less than 230 min in duration were likely to underestimate overall duration, and sessions of greater than 230 min were more likely to overestimate durations. Wirth et al. reported that MTS with longer intervals produced the systematic biases more often than when shorter intervals were used, however there do not appear to have been any studies showing the effect of dimensions of behaviour on MTS biases.

**Interval duration.** As found by Powell et al. (1977), the data from Study 2 showed that MTS with intervals of $\geq 300$ s did not produce representative samples and that, in general, shorter intervals produced more representative samples. As previously mentioned, several studies have suggested that choosing an MTS interval duration shorter than the lower limit of the bout duration will produce representative samples (e.g., Green, McCoy, Burns, &
Smith, 1982; Ary & Suen, 1983). Ary and Suen suggested that for bout durations shorter than the shortest practical MTS interval duration, a frequency count may be a more suitable measure, as duration may not be a dimension of interest.

It is important to note that when using very long MTS interval durations (e.g., 3,600 s or one per hour) there are a limited number of possible obtained durations. For example, conducting 3,600 s MTS over 8 hrs will yield an obtained duration of 0%, 12.5%, 25%, 37.5%, 50%, 62.5%, 75%, 87.5% or 100%. Therefore, samples of behaviours of overall durations falling between the possible obtained durations are unable to fall within stringent accuracy limits, simply due to the impossibility of the mathematics (e.g., a behaviour of 20% overall duration cannot be sampled within ± 4% or less absolute accuracy limits). Although observations conducted over a long session duration such as a day have been suggested to be labour-intensive (Mansell, Jenkins, Felce, & de Kock, 1984), a way to make long observation sessions less labour-intensive is to increase MTS interval duration, allowing the practitioner to measure other behaviours or perform other tasks between observations. However, the practicality of conducting MTS with longer intervals must be weighed against the likely representativeness of the resultant data, particularly because the results of Study 2 show that intervals of ≥ 300 s do not produce representative samples. Few studies have evaluated the practicality of conducting MTS with varying interval durations (variable-interval MTS) across a range of session durations.

**Session duration.** Most studies evaluating the representativeness of samples obtained across various session durations have only evaluated shorter session durations (e.g., 60 min; Devine, Rapp, Testa, Henrickson, & Schnerch, 2011; Kearns, Edwards, & Tingstrom, 1990; Tiger et al., 2013). An exception is Mudford, Beale, and Singh (1990), who evaluated the representativeness of extracted continuous samples of up to 135 min in duration. Although the literature shows that increasing session duration increases
representativeness, no studies have been able to show required session duration relative to the sampling method used and the dimensions of behaviour. Wirth et al. (2014) compared the representativeness of 1-hr, 4-hr, and 8-hr samples with a range of MTS interval durations, showing that longer observation sessions produced more representative data. Their data also showed that 1-hr observation sessions with short MTS intervals (e.g., 30 s) produced more representative samples than 8-hr observations with longer MTS intervals (e.g., 450 s).

Additional to the general findings reported by Wirth et al., Study 2 shows the session duration at which each MTS interval duration produces representative data for behaviours of varying overall and bout duration. The advantage of these data is that practitioners are able to select a sampling method based on the session duration that is likely to be practical in the setting in which they are observing behaviour.

LIMITATIONS

**Practical session duration.** In order to evaluate the representativeness of the samples extracted by the various methods in Study 2, some rules had to be developed. For example, the data from the review of duration measures reported in the *Journal of Applied Behavior Analysis* included the recording of the maximum and minimum session durations used by the studies. Of the 10 studies that used MTS, five conducted 5- or 10-min sessions, and five conducted 20- or 30-min sessions. MTS interval durations ranged from 10 s to 300 s (5 min), with the longer interval durations used in the longer session durations. Devine et al., (2011) suggested that session durations of ≤ 60 min were likely to be practical in applied settings. Their suggestion, in combination with a review of the sessions durations used in studies of representativeness, resulted in the selection of 60-min session duration as the maximum duration likely to be practical in applied settings. As previously discussed, practical session duration may vary with the purposes of data collection, and 60 min may
have been too stringent as a limit. However, the purpose of selecting 60 min as a maximum practical session duration was to provide a criterion from which representativeness could be evaluated.

**Accuracy limits.** Another criterion set in Study 2 was that 95% of the 100 simulated samples were required to fall within each selected absolute accuracy limit for the sampling method to be considered to be representative. As multiple simulations were conducted, and the resultant obtained percentages were variable, a rule was required that would determine the likelihood of each sample method to produce samples representative within designated accuracy limits. Expecting all 100 simulated observation samples to fall within a designated accuracy limit was too stringent, and therefore the 95% limit provided a measure of accuracy of the claim that the sampling method produced representative samples. The 95% limits were therefore akin to the principle of confidence intervals, which are calculated based on the mean of the data and provide a measure of accuracy of the mean (Cumming & Finch, 2005). The 95% limits in Study 2 should be interpreted as a measure of the probability that a sample conducted with a particular method, for behaviour of specified duration and bout duration. The limits indicate that 95% of samples conducted using that particular sampling method will fall within the specified accuracy limits.

**Error.** The decision was made to eliminate relative error from Study 2 analyses, based on assumptions of what may be practical or reasonable accuracy limits for behaviours of varying duration and the finding that many relative accuracy limits that could be practical overlapped with absolute accuracy limits. Tiger et al. (2013) suggested that relative error for low-duration behaviours results in very stringent limits, and therefore that absolute error may be more desirable. The advantages of absolute error are that low-duration behaviours are not required to be measured within impractically small limits, and that absolute error is easier to conceptualise and calculate for practitioners. However, it is important to note that the
maximum desirable absolute error is likely to be different for low- and high-duration behaviours. For example, absolute accuracy limits of ± 10% may still produce data sufficiently representative of behaviours occurring with 50% overall duration, but would be too generous for behaviours occurring with 5% overall duration. It is unlikely that obtained percentage of zero (i.e., behaviour is not detected) would be desirable for behaviours of 5% and less overall duration, despite the lower absolute accuracy limits falling on 0%. The categorisation of samples of $P_0 = 0$ as unrepresentative in Study 2 contributed largely to the conclusion that behaviours of 5% or less overall duration are not sampled representatively with MTS. It was decided for the analyses of Study 2 data that a maximum absolute accuracy of ± 5% and ± 10% would be desirable for behaviours of 5% or less overall duration and 10% or greater overall duration, respectively. One of the advantages of the decision rules produced by Study 2 is the ability to select acceptable error.

Several authors have suggested that acceptable error is likely to be affected by the purpose of data collection (Taylor, Skourides, & Alvero, 2012; Tiger et al., 2013). Taylor et al. suggested that being able to predict that direction of errors a sampling method may produce (i.e., under- and overestimations of duration) may help to choose a sampling method depending on the severity or importance of the behaviour. For example they suggested that for desirable behaviours such as wearing safety equipment, overestimations would be undesirable. Therefore, it could be suggested that for behaviours that are to be decreased, underestimations of duration are undesirable, and for behaviours that are to be increased, overestimations of duration are undesirable. Despite the implications of being able to predict the direction of error of a sampling method, there is little research or discussion of the size of acceptable error in samples collected for various purposes.

Barrios (1988) identified the following purposes for data collection: 1. to screen a person for suitability of treatment; 2. assessment of problem behaviour; 3. treatment
selection; and, 4. treatment evaluation. It is likely that it would be desirable to have constant acceptable error in the assessment of problem behaviour and the evaluation of the effectiveness of the subsequent treatment in order to demonstrate functional relation (i.e., that the treatment was responsible for behaviour change). However, because the overall duration of behaviour affects the representativeness of the sampling method, this poses a problem as to how to select a measurement system that will measure behaviours of different durations (i.e., pre- and post-intervention) with the same degree of error. If the behaviour change is an increase in duration, there is little issue because higher-duration behaviours are sampled more representatively than lower-duration behaviours, and if the measurement system selected produces representative samples of the low-duration, pre-intervention behaviour, it is likely to continue to produce representative samples as the behaviour increases in duration. However, decreasing behaviours pose more of a problem. The ability of a sampling method to detect changes in behaviour may be more important than the representativeness of each data point. Some studies such as by Devine et al. (2011) have evaluated the ability of MTS to detect changes in duration, finding that changes were more likely to be detected with longer sessions and shorter MTS intervals. However, Devine et al. only evaluated overall durations of 25% and above. Further study of the most representative sampling methods for changing duration is warranted, particularly if the data can result in data-based decision rules, rather than general guidelines for session duration and MTS interval duration as suggested by studies such as Devine et al.

Another limitation of the decision rules is the requirement of estimations of overall and bout duration. The possible methods that could be used to obtain the estimates are those suggested in the work sampling literature. Either the practitioner can use informant methods to obtain an estimate (i.e., ask parents, teachers, carers), which may produce inaccurate estimates, or conduct a preliminary observation prior to selecting a sampling method. Rapp,
Colby-Dirksen, Michalski, Carroll, and Lindenberg (2008) suggested that 10 min of continuous recording should be conducted prior to decisions regarding desired behaviour change. The data collected could also be used to estimate overall and bout duration. Further study on the amount of data required for an accurate estimate would be of benefit, because the accuracy of the continuous recording is likely to be influenced by the variables that have been demonstrated to affect MTS. In Study 1, it was found that inaccurate preliminary estimates could be corrected for in the recalculation of the required number of observations by the work sampling equations. However, the decision rules in Study 2 do not include recalculation in the method, although the practicality and effect on representativeness re-selecting variables based on the current estimate of behaviour could be studied further.

**STRENGTHS**

One of the strengths of Study 2 is the ability to demonstrate the utility of the decision rules in measuring naturalistic data. Other studies have either conducted simulations (e.g., Wirth et al., 2014) or measured behaviours in applied settings (e.g., Murphy & Goodall, 1980), but few have conducted both. An exception was Powell (1984b), who developed a modified MTS procedure using computer simulations, and then conducted an accuracy analysis using the method to sample visual-fixation behaviour in a young boy in a home setting. Conducting both simulated analyses and analyses in applied settings is beneficial because the combination of analyses allow for systematic manipulation and control over variables, and the assessment of clinical utility. Simulations alone may have little generality to applied settings (Wirth et al.), but applied studies only allow for the effect of only a small range of variables on representativeness to be studied. The results of Study 2 were also useful for suggesting why simulation studies can have limited generality. In particular, the differences in simulated and measured bout durations and interbout intervals suggest that
distribution should be more thoroughly studied in future research (i.e., both typical or likely patterns of behavioural distribution, and their effect on representativeness).

DATA-BASED DECISION RULES

The desired final outcome of Study 2 was to develop a set of data-based decision rules that could be used by practitioners to select a representative sampling method. Upon reviewing studies that assessed the effect of behaviour and sampling variables on the representativeness of samples, Rojahn and Kanoy (1985) suggested that the research provided general, difficult to combine rules regarding selecting sampling method. In addition, Wirth et al. (2014) acknowledged the need for a data-based tool for practitioners. Study 2 was able to produce data-based decision rules in the form of 3-D graphs from which decisions could be made regarding sampling method.

The coloured 3-D mesh plots provide general guidance for selecting a sampling method, but the ability to rotate them is important because different parts of the plot become easier to read accurately at different angles. However, the plots were generated in SigmaPlot 11.0 (Systat Software Inc.), to which few practitioners are likely to have access. The 3-D mesh plots produced using MATLAB® R2011b (Mathworks) provided an alternative from which values could be determined more accurately. The mesh plots resolved the issue that SigmaPlot 11.0 extrapolated the data beyond the plotted values (which suggested that longer observation sessions may produce representative samples for some combinations of variables that did not produce representative samples). A further step towards ensuring usability of the decision rules is the development of a computer program, preferably in the ubiquitous Microsoft Excel, that will permit a practitioner to select estimates of overall and bout duration, a desired accuracy limit and MTS interval duration in order for the program to produce the required session duration. Alternatively, the maximum possible session duration
could be selected in order to determine the accuracy limits for MTS of different intervals. The use of drop-down menus could be used to select the parameters, making the program quick and easy to use. Rojahn and Kanoy (1985) suggested the development of software to assist practitioners to choose measurement systems based on variables such as error or dimensions of behaviour. Easy-to-use software would be more useful for practitioners than complex graphs and would not require large numbers of graphs depicting the combinations of different variables (such as those presented by Wirth et al., 2014).

CONCLUSION

Study 2 achieved the aim of developing data-based decision rules from which practitioners could select sampling methods. The decision rules provide an approach to choosing a sampling method that is advantageous over ad hoc methods such as choosing based on the convenience of session time (Mansell, 1985) or on general rules of thumb such as ‘longer sessions produce more representative data’ (e.g., Devine et al., 2011). Data-based methods should allow for more precise prediction of the parameters that will produce a representative sample (Mazur, 2005), and are consistent with the empirical nature of decision-making in applied behaviour analysis. The decision rules were demonstrated to produce representative samples of some of the behaviours for which criterion records were used from Study 1, and the factors that limited the representativeness of the selected sampling methods were identified. Specifically, the decision rules did not produce representative samples when interbout intervals were more variable than the interbout intervals used in the simulations from which the decision rules were developed. The simulations of various combinations of dimensions of behaviour and sampling parameters corroborated the literature on the effects of each variable on representativeness, and extended the application of the literature to data obtained in applied settings.
Within the small body of research on the representativeness of direct observation samples of behaviour, there are few studies that have offered a data-based procedure for selecting representative measurement systems, and even fewer that have tested the efficacy of such systems with data from applied settings. The use of mathematical models may be advantageous over more general rules (e.g., shorter MTS intervals are better than long intervals) to choose measurement systems as they may allow more precise selection of representative samples (Mazur, 2006). The purpose of both Studies 1 and 2 was to evaluate a data-based procedure for selecting a representative sampling method.

Study 1 adapted work sampling methods from the field of time-and-motion study in industry. The results showed that the work sampling equation using relative accuracy produced representative samples, but that the effort required to obtain the samples was high. The work sampling equation using absolute accuracy was less likely to produce representative samples, but the number of required MTS observations was low (i.e., low effort). The results illustrated a necessary compromise between representativeness and effort. In regards to using work sampling equations to select the length of continuous observation sessions, only lower-duration behaviours (occurring for \( \leq 5\% \) of the week) were sampled representatively using the relative accuracy equation and, therefore, high effort was required. The work sampling equation using absolute accuracy did not produce representative samples when observations were conducted using continuous sampling.

The work sampling equations were not useful for selecting the number of seconds of continuous sampling sessions, and either required impractical numbers of observations to
produce a representative sample (relative accuracy), or small numbers of observations that did not produce representative samples (absolute accuracy). Therefore, work sampling may not be useful as a data-based procedure for selecting the number of required observations, but could be used as an empirical method to determine when MTS is impractical. Similarly, work sampling could be used to determine the likely representativeness of samples conducted using specified numbers of observations.

Study 2 showed that data-based decision rules for selecting a representative sampling method could be developed for a limited range of behaviours with specific values of response bout duration and overall duration. Study 2 was able to show that some behavioural dimensions (e.g., overall duration of \( \leq 5\% \), bout durations of \( \geq 90 \) s) cannot be sampled representatively using momentary time sampling (MTS). In addition, Study 2 was able to show that MTS with intervals \( \geq 300 \) s do not produce representative samples. Study 2 also showed that the generality of the decision rules for behaviours that were shown to be sampled representatively within the selected accuracy limits was limited, due to the differences in the distributions of simulated and naturalistic data.

Work sampling and the data-based decision rules both showed that obtaining representative data required effortful sampling. The two methods may be useful when effortful sampling is feasible over a small number of days (e.g., to obtain baseline data on overall daily durations of behaviours, or to evaluate generalization post-intervention). Other purposes of data collection such as treatment evaluation and functional analyses, in which the environment can be controlled more tightly, may best be suited to other methods of recording such as continuous recording. It is important to note that although both studies only evaluated free-operant behaviours across a long time-of-interest, there are several occasions when it may be desirable to sample behaviour across a long time-of-interest: a work day, a school day, or a particular work shift in a residential service.
DETERMINING REPRESENTATIVENESS

Prior to determining whether data are representative, a practitioner or researcher must determine how representativeness will be measured; relative or absolute accuracy must be selected. The relative accuracy of obtained data is calculated by the relative percentage difference between obtained and true values \([P_O/P_T] \times 100\%\); e.g., Gardenier, MacDonald, & Green, 2004). The absolute accuracy of obtained data is determined by calculating the absolute difference between obtained and true values (e.g., Tiger et al., 2013).

Study 1 evaluated relative and absolute accuracy of ± 5%. Larger error may be acceptable for some data collections, particularly for behaviours occurring for ≤ 10% of the week. Study 2 disregarded relative accuracy and evaluated absolute accuracy of ± 5% for low-duration behaviours (≤ 5% overall duration), and absolute accuracy of ± 10% for intermediate- and high-duration behaviours. However, the accuracy limits were selected somewhat arbitrarily in Study 2, and there may be settings in which less or more stringent absolute accuracy or relative accuracy is required. For example, relative accuracy (i.e., most stringent) may be desirable when measuring dangerous, but low-duration behaviours (≤ 5% overall duration). The lack of uniformity in measuring representativeness across the literature can make general conclusions regarding representativeness difficult. However, the subjective nature of representativeness, as influenced by the situation in which data are being collected, precludes a single standard of representativeness.

UTILITY OF DATA-BASED METHODS

There are two major considerations for the use of data-based procedures for selecting sampling methods: 1. the assumptions or data on which they are based are generalizable to applied settings, and; 2. that the method is practical to use.
**Limitations of assumptions.** As illustrated in Study 2, developing models or testing a method requires making assumptions about dimensions of behaviour, and selecting limits to the range of values of dimensions of behaviours and sampling methods that are modelled. For example, Wirth, Slaven, and Taylor (2014) acknowledged that in their simulations, bout duration and overall duration were confounded. A limitation to models for selecting sampling interval lengths such as the model proposed by Ary and Suen (1986) is the assumption regarding dimensions of behaviour (i.e., that bout durations are normally distributed; Quera, 1990). The data from Study 1, and the subsequent variation in the representativeness of samples selected using the Study 2 decision rules, show that bout durations are unlikely to be normally distributed.

Quera (1990) proposed an alternative model, based on Weibull and exponential distributions, and demonstrated that applying a post hoc correction to the model proposed by Ary and Suen (1986) produced representative samples of duration. Although no statistical analyses of the distribution of bout analyses in Study 1 were conducted, the distributions appear likely to be exponential, characterised by Quera as more likely to contain short than long bout durations, and with bout length independent of the point in time in which the bout commences. However, the models and methods proposed by Ary and Suen and Quera contain complex equations unlikely to be useful for practitioners, and all their analyses involved simulated data. Developing models based on simulated data is useful in order to evaluate systematically the effect of varying the values of the dimensions of behaviour and sampling methods, however the models may not be generalizable to naturalistic data.

Distribution of bouts of responding is a dimension of behaviour that was not varied in the simulations used to develop the decision rules in Study 2. Wirth et al. (2014) acknowledged that only evaluating randomly-distributed behaviour (as in Study 2) was limiting, but that the effect of distribution on representativeness was unknown. The results of
Study 2 do not show the effect of different distributions on representativeness in a systematic way. However, Study 2 results do indicate that as behaviour becomes less randomly distributed (i.e., interbout intervals are more variable), representativeness decreases. The distribution of behaviour is idiosyncratic to an individual's environment, and unlikely ever to be considered random. Rather, it is likely to be preceded by more or less consistent environmental antecedents and followed by consistent consequences that could be identified and measured. Therefore, perhaps the conclusion to be drawn is that attempting to sample behaviour in the ways evaluated in Studies 1 and 2 (i.e., across a whole or part of a day with discontinuous sampling) requires further analysis and quantification of the dimension of distribution. It is possible that the range of possible distributions is too broad and unable to be quantified, which may suggest that sampling in the ways evaluated in Studies 1 and 2 may only be appropriate for a narrow range of behaviours and settings. The results of Studies 1 and 2 suggest that behaviour is unlikely to be randomly distributed in naturalistic settings, and that behaviour that is not randomly distributed is not likely to be sampled representatively using discontinuous sampling.

**Practicality of data-based methods.** The utility of work sampling equations for practitioners may be improved by the use of a spread sheet similar in nature to the one developed by Fisher, Kelley, and Lomas (2003) to assist training in visual inspection. A spread sheet was developed as part of the data analysis for Study 1, which could be used by practitioners. A practitioner can enter an estimated initial $P_o$ value and desired accuracy, and be provided with the required number of observations calculated by whichever work sampling equation has been selected. Results from each day of observation will populate a graph in the spread sheet when entered into the designated boxes. The use of such a spreadsheet may obviate some of the effort and difficulty associated with calculating and recalculating the required number of observations. A potential use of the spread sheet is for
practitioners to determine whether the required number of observations is impractically high, and therefore discard MTS as an option for their data collection.

The 3-D plots developed in Study 2 provide a way of choosing sampling method through visual inspection of figures representing the decision rules. However, a spread sheet would assist practitioners to determine accurately either session length or predicted accuracy when estimates of dimensions of behaviour and practical sampling parameter limits are entered. Simple, efficient methods that require little response effort (e.g., spreadsheets) may increase the likelihood of practitioners using them, and promote the use of data-based procedures for choosing measurement systems.

SUGGESTIONS FOR FUTURE RESEARCH

Despite an increase in the reported use of continuous recording of behaviour (Mudford, Taylor, & Martin, 2009), the results of the analysis of reported durations in the *Journal of Applied Behavior Analysis* suggested that discontinuous sampling remains in common usage. However, as the results of Study 2 demonstrated that MTS does not produce representative samples for some values of behavioural dimensions (overall duration and bout duration), the representativeness of other recording methods should be further explored. For example, aside from Mudford, Beale, and Singh (1990) and Tiger et al. (2013) few studies have evaluated the representativeness of continuous samples of behaviour. The results of Study 1 corroborated the results of Mudford et al. and Tiger et al. by showing that continuous recording across shorter sessions (≤ 60 min) did not produce representative samples. Therefore, if continuous recording is used, consideration should be given to observing the whole time-of-interest. Further research could compare the representativeness of data collected by conducting one or two whole time-of-interest continuous recording sessions with the results of multiple, shorter, continuous recording sessions. Such research could provide
support for the assertion that a data-based procedure for selecting continuous observation duration is unlikely to be possible due to the impractical durations of continuous observations required for representative samples.

The conclusions drawn from Studies 1 and 2 are limited to specific values of behavioural dimensions (overall duration and bout duration) and sampling methods (MTS with a limited range of interval lengths) across a day or week as the time-of-interest. Further research could vary the values of the dimensions of behaviour, or determine the effect of dimensions of behaviour not evaluated in Study 2 on representativeness. Specifically, the results of Study 2 suggest that further research on quantifying distribution and evaluating the effects of different distributions on representativeness is warranted. Furthermore, the data presented by Mudford, Locke, and Jeffrey (2011) could be used to inform simulations of varying rates of behaviour, as rate was not studied in the current research.

The finding that low-duration behaviours were not sampled representatively, and the subsequent suggestions for alternative sampling methods, indicates a direction for future research (i.e., automated measurement or frequency recording). Of concern is that in the analysis of the range of durations reported in the Journal of Applied Behavior Analysis, approximately 25% of the data points collected by MTS depicted behaviour of \( \leq 5\% \) duration. A critical review of the literature in which MTS was used to record low-duration behaviours could question the validity of some of the reported results. However, it should be noted that the current research explored the representativeness of single data points, whereas in the repeated measures experimental designs common in applied behaviour analysis, behaviour is measured multiple times in each phase of a study or treatment. That is, while there may be more error than desirable in any single measurement, the practice of replicating measures several times strengthens the claim that a set of data points is representative of true behaviour.
Tiger et al. (2013) suggested more research to develop decision rules for teachers regarding the number and duration of observations that would probably produce representative samples across a range of patterns of trend, stability, and level in the data. They suggested using Generalizability (G) theory. A repeated-measures approach to representativeness (e.g., the representativeness of multiple data points across a baseline) may, in part, resolve the issue of how to representatively measure low-duration behaviours. Believability of the data increases with each subsequent data point, particularly if the data are stable. Therefore, even if low-duration (≤ 5%) behaviour is only detected in three of ten baseline sessions, these data may be sufficient for the purposes of data collection (i.e., as an indication of overall duration across the baseline, rather than the practitioner being concerned with the representativeness of a single data point). Future research may compare the representativeness of a single long observation session (e.g., 60 min) with multiple shorter sessions (e.g., six 10-min sessions).

CONCLUSION

The use of direct observation to make data-based clinical decisions, and the limited opportunities to observe the full time-of-interest makes the obtaining of representative samples important to applied behaviour analysis. Despite a body of research that provides some guidance on choosing a measurement system, there remains the need for a data-based procedure for choosing sampling methods for behaviours for which those sampling methods produce representative samples. A data-based procedure is consistent with the methods of the field, and would increase the believability of the data reported in research and clinical settings. The current research evaluated two potential data-based procedures for choosing sampling methods to measure the duration of behaviour; work sampling, and decision rules derived from systematic simulations. Both studies showed that representativeness will
increase with increased effort in the measurement system, and both studies showed the limitations and potential uses of the data-based system they evaluated. The current research resulted in some clear directions for future research in metrology, from which applied behaviour analysis would benefit. The field should continue to be concerned with issues of representativeness, as believable data are the foundation on which our decisions, claims, and future success lie.
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Absolute accuracy.</strong></td>
<td>Absolute difference between obtained and true values ($P_O - P_T$).</td>
</tr>
<tr>
<td><strong>Accuracy.</strong></td>
<td>Degree to which dimensions of behaviour such as rate and duration are reflected in a sample.</td>
</tr>
<tr>
<td><strong>Accuracy limits.</strong></td>
<td>Upper and lower values on either side of the true value within which a specified percentage of the data points should fall. E.g., when desired accuracy is ± 5%, 95% of data should fall within the accuracy limits.</td>
</tr>
<tr>
<td><strong>Bout duration.</strong></td>
<td>Duration (usually reported in s or min) of a single occurrence of a behavioural or environmental event, measured by recording the time between the onset and offset of the occurrence.</td>
</tr>
<tr>
<td><strong>Continuous recording.</strong></td>
<td>Recording a dimension of behaviour such as rate and duration as each instance occurs in real time or from video recordings.</td>
</tr>
<tr>
<td><strong>Conventional true value.</strong></td>
<td>Value used as a true value for the purposes of the data analysis.</td>
</tr>
<tr>
<td><strong>Dimension.</strong></td>
<td>Characteristic of behaviour such as duration, rate, frequency, or inter-response time.</td>
</tr>
<tr>
<td><strong>Discontinuous recording.</strong></td>
<td>Division of observation sessions into intervals of time, with occurrences of behaviour recorded within or at the end of each interval.</td>
</tr>
<tr>
<td><strong>Distribution.</strong></td>
<td>Description of the locus in time of each behavioural or environmental event across the whole time-of-interest (i.e., pattern).</td>
</tr>
<tr>
<td><strong>Error.</strong></td>
<td>Difference between the true value and the obtained value.</td>
</tr>
<tr>
<td><strong>Interval length.</strong></td>
<td>Length of time between observations in discontinuous recording.</td>
</tr>
</tbody>
</table>
Measurand. Variable to be measured. A behaviour or environmental event in applied behaviour analysis.

Momentary time sampling. Observation session is divided into intervals and a decision is made as to whether a behaviour is occurring or not at the end of each interval.

Nomograph. Diagram of graduated lines, each representing a variable, from which values of an unknown variable can be read when values of the other variables are known. Work sampling equations can be represented in nomographs.

Overall duration. Cumulative duration of all occurrences of a behaviour across a specified period (e.g., time-of-interest), reported in s or min, or in percentage of the day, week, or intervals.

Observation. Single instance of observing a behaviour (such as the instantaneous observation conducted at the end of each interval in momentary time sampling).

Observation session. Period during which behaviour is observed and measured.

Value of dimension. Description of the range of a dimension of behaviour (e.g., describing a behaviour as occurring between 15 and 20 times per hour refers to the values of the rate of the behaviour).

Partial interval recording. Behaviour is recorded in each interval if it has occurred once or more during that interval.

Representativeness. Degree to which the data obtained from a sample observation session reflects overall dimensions of behaviour, across all opportunities to emit the behaviour (the full time of interest).

Relative accuracy. Relative percentage difference between obtained and true values ($[P_o/P_T] \times 100\%$).

Reliability. Measure of the stability of the measurement system; that the data collected over successive samples will be consistent if behaviour is consistent.

Response. Single occurrence of a behaviour.
<table>
<thead>
<tr>
<th><strong>Sample.</strong></th>
<th>Obtained measure of a dimension of behaviour.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Snap observation.</strong></td>
<td>Terminology used in work sampling to describe a single, instantaneous momentary time sample observation.</td>
</tr>
<tr>
<td><strong>Systematic error.</strong></td>
<td>Errors that are consistent across all observation sessions conducted using a particular measurement system (e.g., duration is underestimated in all samples).</td>
</tr>
<tr>
<td><strong>Time-of-interest.</strong></td>
<td>Period of time to be sampled by observation sessions.</td>
</tr>
<tr>
<td><strong>True value.</strong></td>
<td>Quantitative measures of a dimension of a behaviour that reflect the true state of that dimension.</td>
</tr>
<tr>
<td><strong>Validity.</strong></td>
<td>Degree to which a measurement system measures what it is designed to measure.</td>
</tr>
<tr>
<td><strong>Whole interval recording.</strong></td>
<td>Behaviour is scored as occurring if it occurred throughout the whole interval.</td>
</tr>
<tr>
<td><strong>Work sampling.</strong></td>
<td>Quantitative method of producing cost-efficiently representative samples of how people are spending their time. Based on statistical sampling techniques and on the laws of probability, and from the field of time-and-motion study.</td>
</tr>
</tbody>
</table>
REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


MATLAB® (Version R2011b) [Computer software]. Natick, MA: The Mathworks, Inc.


REFERENCES


REFERENCES


REFERENCES


Appendix 1

TEACHER ESTIMATES OF $P_O$
How do you think time was spent today? – Participant 1

Participant 1: 09-09-11

Please estimate the percentage of time you think the student spent doing each of these things TODAY. Percentages do not have to add up to 100% as students can be doing multiple things at once (for example, engaging in problem behaviour such as vocal stereotypy whilst also being off task, although these things do not necessarily always occur together).

**Example: What percentage of time today did the student interact with peers?**

What percentage of time do you think the student spends toe-walking?

What percentage of time do you think the student spends grabbing?

What percentage of time do you think the student spends on-task?

What percentage of time do you think the student spends happy (smiling, laughing, giggling)?

What percentage of time do you think the student spend unhappy (crying, wailing, frowning)?

What percentage of time do you think the student receives one-on-one attention for?

What percentage of time do you think adults spend asking him to do specific things (stop, sit down etc)?

**What is the best time of the day / week to observe....**

<table>
<thead>
<tr>
<th>Activity</th>
<th>Day</th>
<th>Week</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toe-walking</td>
<td>Anytime</td>
<td>Anytime</td>
</tr>
<tr>
<td>Grabbing</td>
<td>Monday, 1st session of day</td>
<td></td>
</tr>
<tr>
<td>On-task</td>
<td>Middle session, any day</td>
<td></td>
</tr>
<tr>
<td>Happy</td>
<td>Outside, any day</td>
<td></td>
</tr>
<tr>
<td>Unhappy</td>
<td>Class time, any day</td>
<td></td>
</tr>
</tbody>
</table>

*Teacher’s response was 60 to 70%, average taken for analysis
How do you think time was spent today? – Participant 2

Date: 30-09-11

Please estimate the percentage of time you think the student spent doing each of these things TODAY. Percentages do not have to add up to 100% as students can be doing multiple things at once (for example, engaging in problem behaviour such as vocal stereotypy whilst also being off task, although these things do not necessarily always occur together).

**Example: What percentage of time today did the student interact with peers?**

<table>
<thead>
<tr>
<th>Time Activity</th>
<th>Expected Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tantrum</td>
<td>25%</td>
</tr>
<tr>
<td>Grabbing</td>
<td>25%</td>
</tr>
<tr>
<td>On-task</td>
<td>20%</td>
</tr>
<tr>
<td>Happy</td>
<td>55%</td>
</tr>
<tr>
<td>Unhappy</td>
<td>20%</td>
</tr>
<tr>
<td>One-on-one</td>
<td>80%</td>
</tr>
</tbody>
</table>

**Day**

- Tantrum: Before going to respite care
- Grabbing: Non-preferred activities
- On-task: Smartboard
- Happy: During Smartboard / drawing
- Unhappy: Any time

*Teacher’s response was 50 to 60%, average taken for analysis*
How do you think time was spent today? – Participant 3

Date: 08-12-11

Please estimate the percentage of time you think the student spent doing each of these things TODAY. Percentages do not have to add up to 100% as students can be doing multiple things at once (for example, engaging in problem behaviour such as vocal stereotypy whilst also being off task, although these things do not necessarily always occur together).

Example: What percentage of time today did the student interact with peers? [13%]

What percentage of time do you think the student spends in a running? [15%] *

What percentage of time do you think the student spends stereotypy (e.g., finger picking, head in jumper)? [70%]

What percentage of time do you think the student spends on-task? [30%]

What percentage of time do you think the student spends happy (smiling, laughing, giggling)? [90%]

What percentage of time do you think the student receives one-on-one attention for? [40%] *

What percentage of time do you think adults spend asking him to do specific things (stop, sit down etc.)? [20%]

What is the best time of the day / week to observe….

<table>
<thead>
<tr>
<th>Activity</th>
<th>Day</th>
<th>Week</th>
</tr>
</thead>
<tbody>
<tr>
<td>Running</td>
<td>Any day, first thing in morning</td>
<td></td>
</tr>
<tr>
<td>Stereotypy</td>
<td>Any time</td>
<td>Any time</td>
</tr>
<tr>
<td>On-task</td>
<td>Work experience</td>
<td>Morning</td>
</tr>
<tr>
<td>Happy</td>
<td>Swimming</td>
<td>Any time</td>
</tr>
</tbody>
</table>

*Teacher’s response was 10 to 20% and 30 to 50% for running and attention respectively, average taken for analysis
Appendix 2

DURATION ANALYSIS
### Appendix 2.

**Articles from the Journal of Applied Behavior Analysis (2008 to 2012) and Variables Measured for Duration Analysis.**

<table>
<thead>
<tr>
<th>Article</th>
<th>Issue</th>
<th>Measure</th>
<th>Interval dur.</th>
<th>Session dur.</th>
<th>Behaviour topography</th>
<th>Behaviour change</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athens, Vollmer, Sloman &amp; St. Pipkin (2008)</td>
<td>2</td>
<td>Continuous</td>
<td>-</td>
<td>5 min</td>
<td>Vocal stereotypy</td>
<td>Decrease</td>
<td>Home</td>
</tr>
<tr>
<td>Austin &amp; Soeda (2008)</td>
<td>2</td>
<td>PIR</td>
<td>10 s</td>
<td>40 min</td>
<td>Off-task behaviour</td>
<td>Decrease</td>
<td>Classroom</td>
</tr>
<tr>
<td>Betz, Higbee, &amp; Reagon (2008)</td>
<td>2</td>
<td>MTS</td>
<td>20 s</td>
<td>20 min</td>
<td>Peer engagement</td>
<td>Increase</td>
<td>Classroom</td>
</tr>
<tr>
<td>Brenske, Rudrud, Schulze, &amp; Rapp (2008)</td>
<td>2</td>
<td>MTS</td>
<td>60 s</td>
<td>40 min</td>
<td>Activity engagement</td>
<td>Increase</td>
<td>Nursing home</td>
</tr>
<tr>
<td>DeLeon et al (2008)</td>
<td>1</td>
<td>Continuous</td>
<td>-</td>
<td>10 min</td>
<td>Wearing glasses</td>
<td>Increase</td>
<td>Treatment room</td>
</tr>
<tr>
<td>Lang et al. (2008)</td>
<td>3</td>
<td>PIR</td>
<td>10 s</td>
<td>5 min</td>
<td>Head hitting, aggression, elopement, dropping to floor</td>
<td>Functional analysis</td>
<td>Classroom, assessment room</td>
</tr>
<tr>
<td>McKenzie, Smith, Simmons, &amp; Soderlund (2008)</td>
<td>2</td>
<td>Continuous</td>
<td>-</td>
<td>10 min minimum</td>
<td>Eye poking</td>
<td>Decrease</td>
<td>Treatment room, home, canteen</td>
</tr>
<tr>
<td>Roane &amp; Kelley (2008)</td>
<td>3</td>
<td>Continuous</td>
<td>-</td>
<td>5 or 10 min</td>
<td>Foot withdrawals (walking)</td>
<td>Decrease</td>
<td>Clinic</td>
</tr>
<tr>
<td>Roscoe, Carreau, MacDonald, &amp; Pence (2008)</td>
<td>3</td>
<td>MTS</td>
<td>10 s</td>
<td>3 or 5 min</td>
<td>Self-biting, motor stereotypy, tapping, shirt twirling, item engagement</td>
<td>Functional analysis</td>
<td>Clinic</td>
</tr>
<tr>
<td>Study (Year)</td>
<td>Duration</td>
<td>Frequency</td>
<td>Stimulus</td>
<td>Outcome</td>
<td>Methodology</td>
<td>Setting</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>----------</td>
<td>-----------</td>
<td>----------</td>
<td>---------</td>
<td>-------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>Tasky, Rudrud, Schulze, &amp; Rapp (2008)</td>
<td>2</td>
<td>MTS</td>
<td>5 min</td>
<td>30 min</td>
<td>On-task behaviour</td>
<td>Increase</td>
<td>Hospital</td>
</tr>
<tr>
<td>Camp, Iwata, Hamond, &amp; Bloom (2009)</td>
<td>2</td>
<td>PIR</td>
<td>10 s</td>
<td>15 min</td>
<td>Property destruction</td>
<td>Functional analysis</td>
<td>Treatment room</td>
</tr>
<tr>
<td>Delmendo, Borrero, Beauchamp, &amp; Francisco (2009)</td>
<td>3</td>
<td>PIR</td>
<td>10 s</td>
<td>Not specified</td>
<td>Consumption of food</td>
<td>Preference assessment</td>
<td>Treatment room, home</td>
</tr>
<tr>
<td>Dorey, Rosales-Ruiz, Smith, &amp; Lovelace</td>
<td>4</td>
<td>Continuous</td>
<td>-</td>
<td>10 min</td>
<td>Self-injurious behaviour</td>
<td>Functional analysis, decrease</td>
<td>Zoo enclosure</td>
</tr>
<tr>
<td>Gardner, Wacker, &amp; Boelter (2009)</td>
<td>2</td>
<td>PIR</td>
<td>6 s</td>
<td>5 min</td>
<td>Inappropriate behaviour</td>
<td>Functional analysis</td>
<td>Classroom, playground</td>
</tr>
<tr>
<td>Hagopian, Kuhn, &amp; Strother (2009)</td>
<td>4</td>
<td>PIR</td>
<td>2 min</td>
<td>90 to 120 min</td>
<td>Social withdrawal</td>
<td>Decrease</td>
<td>Inpatient unit</td>
</tr>
<tr>
<td>Lang et al. (2009)</td>
<td>2</td>
<td>PIR</td>
<td>10 s</td>
<td>5 min</td>
<td>Screaming, hitting</td>
<td>Functional analysis</td>
<td>Classroom, playground</td>
</tr>
<tr>
<td>Lang et al. (2009)</td>
<td>4</td>
<td>PIR</td>
<td>10 s</td>
<td>10 min</td>
<td>Stereotypy, challenging behaviour, functional play</td>
<td>Motivating operation analysis</td>
<td>Treatment room</td>
</tr>
<tr>
<td>Mace et al. (2009)</td>
<td>1</td>
<td>PIR</td>
<td>10 s</td>
<td>5 min</td>
<td>Activity engagement, aggression</td>
<td>Motivating operation analysis</td>
<td>Home</td>
</tr>
<tr>
<td>Miguel, Clark, Tereshko, &amp; Ahearn (2009)</td>
<td>4</td>
<td>Continuous</td>
<td>-</td>
<td>5 min</td>
<td>Vocal stereotypy</td>
<td>Functional analysis, decrease</td>
<td>Treatment room</td>
</tr>
</tbody>
</table>
| Study Reference | Sample Size | Intermittent 
IR | Duration | Target Behaviour | Analysis Method | Setting |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>O'Reilly et al. (2009)</td>
<td>4</td>
<td>PIR</td>
<td>10 s</td>
<td>10 min</td>
<td>Loud vocalisations, throwing objects</td>
<td>Functional analysis</td>
</tr>
<tr>
<td>Pence, Roscoe, Bourret, &amp; Ahearn (2009)</td>
<td>2</td>
<td>Continuous</td>
<td>-</td>
<td>5 or 10 min</td>
<td>Motor stereotypy</td>
<td>Functional analysis</td>
</tr>
<tr>
<td>Tiger, Fisher, &amp; Bouxsein (2009)</td>
<td>2</td>
<td>PIR</td>
<td>10 s</td>
<td>Not clear – possibly 7 hr</td>
<td>Skin picking</td>
<td>Decrease</td>
</tr>
<tr>
<td>Anderson, Doughty, Doughty, Williams, &amp; Saunders (2010)</td>
<td>2</td>
<td>PIR</td>
<td>5 s</td>
<td>10 min</td>
<td>Stereotypy</td>
<td>Decrease</td>
</tr>
<tr>
<td>Carter (2010)</td>
<td>3</td>
<td>PIR</td>
<td>10 s</td>
<td>5 min</td>
<td>Destructive behaviour and compliance</td>
<td>Decrease, increase</td>
</tr>
<tr>
<td>Dixon, Nastally, &amp; Waterman (2010)</td>
<td>3</td>
<td>PIR</td>
<td>10 s</td>
<td>45 to 60 min</td>
<td>Indices of happiness</td>
<td>Increase</td>
</tr>
<tr>
<td>Fogel, Miltenberger, Graves, &amp; Koehler (2010)</td>
<td>4</td>
<td>Continuous</td>
<td>-</td>
<td>30 min</td>
<td>Physical activity</td>
<td>Increase</td>
</tr>
<tr>
<td>Grauvogel-MacAleese &amp; Wallace (2010)</td>
<td>3</td>
<td>PIR</td>
<td>10 s</td>
<td>5 min</td>
<td>Off-task behaviour</td>
<td>Functional analysis, decrease</td>
</tr>
<tr>
<td>Lang et al. (2010)</td>
<td>1</td>
<td>PIR</td>
<td>10 s</td>
<td>5 and 30 min</td>
<td>Elopement</td>
<td>Functional analysis, decrease</td>
</tr>
<tr>
<td>McGinnis, Houchins-Juárez, McDaniels, &amp; Kennedy (2010)</td>
<td>1</td>
<td>PIR</td>
<td>10 s</td>
<td>60 min</td>
<td>Problem behaviour</td>
<td>Motivating operation analysis</td>
</tr>
<tr>
<td>Study</td>
<td>Duration</td>
<td>Setting</td>
<td>Behaviour</td>
<td>Response</td>
<td>Environment</td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>----------</td>
<td>------------------</td>
<td>------------------------------------</td>
<td>----------</td>
<td>-------------------</td>
<td></td>
</tr>
<tr>
<td>St. Peter Pipkin, Vollmer, &amp; Sloman (2010) – Study 2</td>
<td>5 min</td>
<td>Therapy room</td>
<td>On- and off-task behaviour</td>
<td>Increase, decrease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plavnick, Ferreri, &amp; Maupin (2010)</td>
<td>10 to 15 min</td>
<td>Classroom</td>
<td>Correctly implemented steps of token economy</td>
<td>Increase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ulke-Kurcuoglu &amp; Kircaali-Iftar (2010)</td>
<td>30 min</td>
<td>Clinic</td>
<td>On-task behaviour</td>
<td>Increase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Waller &amp; Higbee (2010)</td>
<td>Not specified</td>
<td>Classroom</td>
<td>Disruptive and appropriate academic behaviours</td>
<td>Decrease, increase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ahrens, Lerman, Kodak, Wordsell, &amp; Keegan (2011)</td>
<td>5 to 30 min</td>
<td>Home</td>
<td>Vocal and motor stereotypy, compliance, appropriate vocalisations</td>
<td>Decrease</td>
<td>Therapy room</td>
<td></td>
</tr>
<tr>
<td>Campbell &amp; Anderson (2011)</td>
<td>15 min</td>
<td>School</td>
<td>Problem behaviour, peer / adult attention, task avoidance</td>
<td>Decrease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Donaldson &amp; Vollmer (2011)</td>
<td>10 min or circle time (average 13.7 min)</td>
<td>Home</td>
<td>Crying</td>
<td>Decrease</td>
<td>School</td>
<td></td>
</tr>
<tr>
<td>Dozier, Iwata, &amp; Wordsell (2011)</td>
<td>5 min</td>
<td>Workplace</td>
<td>Inappropriate sexual behaviour</td>
<td>Decrease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Groskreutz, Groskreutz, &amp; Higbee (2011)</td>
<td>5 min</td>
<td>Classroom</td>
<td>Vocal stereotypy, item interaction</td>
<td>Decrease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heal &amp; Hanley (2011)</td>
<td>4.5 to 5 min</td>
<td>Therapy room</td>
<td>Play</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>N</td>
<td>Delivery Type</td>
<td>Duration</td>
<td>Antecedent</td>
<td>Outcome</td>
<td>Location</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>---</td>
<td>---------------</td>
<td>-----------</td>
<td>------------</td>
<td>---------</td>
<td>------------------</td>
</tr>
<tr>
<td>Kliebert &amp; Tiger (2011)</td>
<td>4</td>
<td>Continuous</td>
<td>5 min</td>
<td>Rumination</td>
<td>Decrease</td>
<td>Empty classroom</td>
</tr>
<tr>
<td>Jennett, Jann, &amp; Hagopian (2011)</td>
<td>4</td>
<td>Continuous</td>
<td>2 to 5 min</td>
<td>Stimulus contact</td>
<td>Competing stimulus assessment</td>
<td>Treatment room</td>
</tr>
<tr>
<td>Kang et al. (2011)</td>
<td>4</td>
<td>PIR</td>
<td>10 s</td>
<td>Problem behaviours</td>
<td>Functional analysis, preference assessment</td>
<td>Empty classroom</td>
</tr>
<tr>
<td>Lanovaz, Sladeczek, &amp; Rapp (2011)</td>
<td>3</td>
<td>Continuous</td>
<td>15 min</td>
<td>Vocal stereotypy</td>
<td>Decrease</td>
<td>Home</td>
</tr>
<tr>
<td>Mace, Pratt, Prager, &amp; Pritchard (2011)</td>
<td>1</td>
<td>PIR</td>
<td>10 s</td>
<td>Vocalisations, aggression, disruption</td>
<td>Decrease</td>
<td>Classroom</td>
</tr>
<tr>
<td>Morrison, Roscoe, &amp; Atwell (2011)</td>
<td>3</td>
<td>MTS</td>
<td>10 s</td>
<td>Self-injury, stereotypy, item engagement</td>
<td>Functional analysis, decrease</td>
<td>Treatment room</td>
</tr>
<tr>
<td>Reed &amp; Martens (2011)</td>
<td>1</td>
<td>MTS</td>
<td>5 s</td>
<td>On-task</td>
<td>Increase</td>
<td>Classroom</td>
</tr>
<tr>
<td>Richling et al. (2011)</td>
<td>2</td>
<td>Continuous</td>
<td>5 to 20 min</td>
<td>Wearing glasses or orthotics</td>
<td>Increase</td>
<td>Home Classroom</td>
</tr>
<tr>
<td>Rispoli et al. (2011)</td>
<td>1</td>
<td>PIR, WIR</td>
<td>10 s, 10 s</td>
<td>Problem behaviour, Academic engagement</td>
<td>Decrease, Increase</td>
<td>Classroom</td>
</tr>
<tr>
<td>Rivas et al. (2011)</td>
<td>2</td>
<td>Continuous</td>
<td>20 min</td>
<td>Crying</td>
<td>Decrease</td>
<td>Treatment room</td>
</tr>
<tr>
<td>Study Authors/Year</td>
<td>Duration</td>
<td>Interventions</td>
<td>Outcome</td>
<td>Setting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>----------</td>
<td>---------------</td>
<td>---------</td>
<td>-----------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rooker, Iwata, Harper, Fahmie, &amp; Camp (2011)</td>
<td>4 PIR 10 s 10 min</td>
<td>Stereotypy</td>
<td>Functional analysis</td>
<td>Treatment room</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schumacher &amp; Rapp (2011)</td>
<td>3 Continuous - 30 min</td>
<td>Vocal stereotypes</td>
<td>Decrease</td>
<td>Home</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sigurdsson, Ring, Needham, Boscoe, &amp; Silverman (2011)</td>
<td>1 WIR 5 s every 40 s 20 min</td>
<td>Leg posture</td>
<td>Increase</td>
<td>Work centre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singh, Ring, Needham, Boscoe, &amp; Silverman (2011)</td>
<td>1 WIR 5 s 5 min</td>
<td>Looking at books</td>
<td>Increase</td>
<td>Classroom</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stocco, Thompson, &amp; Rodriguez (2011)</td>
<td>3 Continuous - 10 min</td>
<td>Engagement with item</td>
<td>Assessment</td>
<td>Treatment room</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thomason-Sassi, Iwata, Neidert, &amp; Roscoe (2011)</td>
<td>1 PIR 10 s 10 min</td>
<td>Problem behaviour</td>
<td>Functional analysis</td>
<td>Treatment room</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allison et al. (2012)</td>
<td>3 Continuous - 5 min</td>
<td>Negative vocalisations</td>
<td>Decrease</td>
<td>Treatment room</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colón, Ahearn, Clark, &amp; Masalsky (2012)</td>
<td>1 MTS Continuous 10 s 5 min</td>
<td>Vocal stereotypes</td>
<td>Functional analysis, decrease</td>
<td>Treatment room</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frank-Crawford et al (2012)</td>
<td>1 PIR Continuous 10 s 2 min</td>
<td>Requests, grabs</td>
<td>Increase</td>
<td>Treatment room</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fritz, Iwata, Rolider, Camp, &amp; Neidert (2012)</td>
<td>1 PIR 10 s 10 min</td>
<td>Stereotypy</td>
<td>Functional analysis, decrease</td>
<td>Treatment room</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Participants</td>
<td>Protocol</td>
<td>Duration</td>
<td>Intervention</td>
<td>Outcomes</td>
<td>Setting</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>--------------</td>
<td>----------</td>
<td>----------</td>
<td>--------------</td>
<td>----------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>Hustyi, Normand, Larson, &amp; Morley (2012)</td>
<td>2</td>
<td>PIR</td>
<td>5 s</td>
<td>5 min</td>
<td>Physical activity Assessment</td>
<td>Playground</td>
</tr>
<tr>
<td>Kadey &amp; Roane (2012)</td>
<td>2</td>
<td>Continuous</td>
<td>-</td>
<td>5 min</td>
<td>Physical activity</td>
<td>Home</td>
</tr>
<tr>
<td>Lanojaz, Rapp, &amp; Ferguson (2012)</td>
<td>4</td>
<td>Continuous</td>
<td>-</td>
<td>5 to 10 min</td>
<td>Vocal stereotypy Decrease</td>
<td>Community centre, home</td>
</tr>
<tr>
<td>Long, Wilder, Betz, &amp; Dutta (2012)</td>
<td>4</td>
<td>Continuous</td>
<td>-</td>
<td>15 min</td>
<td>On-task (computer game) Increase</td>
<td>Simulated office</td>
</tr>
<tr>
<td>Love, Miguel, Fernand, &amp; La Brie (2012)</td>
<td>3</td>
<td>Continuous</td>
<td>-</td>
<td>5 min</td>
<td>Vocal stereotypy Functional analysis, decrease</td>
<td>Treatment room</td>
</tr>
<tr>
<td>Ortega, Iwata, Nogales-González, &amp; Frades (2012)</td>
<td>4</td>
<td>PIR</td>
<td>30 s</td>
<td>2 to 5 min</td>
<td>Playing keyboard Increase</td>
<td>Senior day centre</td>
</tr>
<tr>
<td>Roantree &amp; Kennedy (2012)</td>
<td>3</td>
<td>MTS</td>
<td>5 s</td>
<td>10 min</td>
<td>Social behaviour Functional analysis</td>
<td>Classroom</td>
</tr>
<tr>
<td>Rodriguez, Thompson, Schlichenmeyer, &amp; Stocco (2012)</td>
<td>1</td>
<td>Continuous</td>
<td>-</td>
<td>10 min</td>
<td>Ordering Functional analysis, decrease</td>
<td>Treatment room, living room</td>
</tr>
<tr>
<td>Saylor, Sidener, Reeve, Fetherston, &amp; Progar (2012)</td>
<td>1</td>
<td>MTS</td>
<td>30 s</td>
<td>10 min</td>
<td>Vocal stereotypy Decrease</td>
<td>Home</td>
</tr>
<tr>
<td>Study</td>
<td>N</td>
<td>Type</td>
<td>Duration</td>
<td>Intervention</td>
<td>Goal</td>
<td>Setting</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----</td>
<td>----------</td>
<td>----------</td>
<td>-------------------------------</td>
<td>-------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>Shayne, Fogel, Miltenberger, &amp; Koehler (2012)</td>
<td>1</td>
<td>Continuous</td>
<td>30 min</td>
<td>Physical activity</td>
<td>Increase</td>
<td>Physical education class</td>
</tr>
<tr>
<td>Smith, Smith, Dracoblly, &amp; Pace (2012)</td>
<td>4</td>
<td>PIR</td>
<td>10 s</td>
<td>Disruptive vocalisations, self-injury</td>
<td>Functional analysis</td>
<td>Group home, vocational setting</td>
</tr>
<tr>
<td>Wilson, Iwata, &amp; Bloom (2012)</td>
<td>4</td>
<td>PIR</td>
<td>15 min</td>
<td>Self-injury</td>
<td>Decrease</td>
<td>Group home, vocational setting</td>
</tr>
</tbody>
</table>
Appendix 3

MATLAB® CODE
Example: Overall duration 1% (288 s total), bout duration 1 s to 10 s, 58 events

```
x = randfixedsum(58,1,288,1,10);
tp = .5; tm = -.5;
for k = 1:53
    t0 = (tp+tm)/2;
du = round(x+t0); % Round with offset
e = sum(du)-288;
if e > 0, tp = t0;
elseif e < 0, tm = t0;
else break % Break out when sum is correct
end
end
y = randfixedsum(59,1,28512,1,28512);
tp = .5; tm = -.5;
for k = 1:53
    t0 = (tp+tm)/2;
on = round(y+t0); % Round with offset
e = sum(on)-28512;
if e > 0, tp = t0;
elseif e < 0, tm = t0;
else break % Break out when sum is correct
end
end
h=[0]
dura=[h;du]
b = cumsum(on)
f=cumsum(dura)
onset=b+f
st=onset(1:58)
duration=dura(2:59)
m=cumsum(duration)
t = m - duration + 1;
s = zeros(1,m(end));
s(t) = 1;
ii = cumsum(s);
```
out = (1:m(end)) - t(ii) + st(ii);
sort(out)'
out=unique(out)
xlswrite('1percentsmallbout',sort(out)', 'Sheet1','C1:C290')
Example of MTS sampling (10-s intervals)

t=10:10:600
c=intersect(t,out)
o=length(c)
xlswrite('10percentsmallboutMTS',c, 'Sheet1', 'D1')

t=10:10:1200
c=intersect(t,out)
o=length(c)
xlswrite('10percentsmallboutMTS',c, 'Sheet1', 'D2')

t=10:10:1800
c=intersect(t,out)
o=length(c)
xlswrite('10percentsmallboutMTS',c, 'Sheet1', 'D3')

t=10:10:2400
c=intersect(t,out)
o=length(c)
xlswrite('10percentsmallboutMTS',c, 'Sheet1', 'D4')

...

...

...

...

t=10:10:28800
c=intersect(t,out)
o=length(c)
xlswrite('10percentsmallboutMTS',c, 'Sheet1', 'D48')
Finding number of 100 samples within absolute accuracy limits

```matlab
xlsread('50percentextralargeboutMTS3600', 'Sheet5', 'B1: CW1')
w=ans'
a = find(w>=.49) ;
d = find(w<=.51) ;
c=intersect(a,d) ;
o=length(c)
xlswrite('50percentextralargeboutMTS3600',o,'Sheet6', 'DF1')
xlsread('50percentextralargeboutMTS3600', 'Sheet5', 'B2: CW2')
w=ans'
a = find(w>=.49) ;
d = find(w<=.51) ;
c=intersect(a,d) ;
o=length(c)
xlswrite('50percentextralargeboutMTS3600',o,'Sheet6', 'DF2')
...
...
...
...
xlsread('50percentextralargeboutMTS3600', 'Sheet5', 'B47: CW47')
w=ans'
a = find(w>=.3) ;
d = find(w<=.7) ;
c=intersect(a,d) ;
o=length(c)
xlswrite('50percentextralargeboutMTS3600',o,'Sheet6', 'DY47')
xlsread('50percentextralargeboutMTS3600', 'Sheet5', 'B48: CW48')
w=ans'
a = find(w>=.3) ;
d = find(w<=.7) ;
c=intersect(a,d) ;
o=length(c)
xlswrite('50percentextralargeboutMTS3600',o,'Sheet6', 'DY48')
```
Appendix 4

STUDY 2 ROTATED 3-D MESH PLOTS & 3-D STEM PLOTS
Appendix 4.1. Three-dimensional mesh plot from which required session duration can be determined when acceptable accuracy limits and practical MTS interval duration is selected for behaviours of 10% overall duration and short bout duration (1 s to 10 s). The panels show two different rotations of the plot.
Appendix 4.2. Three-dimensional mesh plot from which required session duration can be determined when acceptable accuracy limits and practical MTS interval duration is selected for behaviours of 90% overall duration and long bout duration (120 s to 300 s). The panels show two different rotations of the plot.
Appendix 4.3. Three-dimensional stem plot from which required session duration can be determined when acceptable accuracy limits and practical MTS interval duration is selected for behaviours of 10% overall duration. Behaviours with short bout duration (1 s to 10 s) are represented in the left plot, and behaviours with medium bout duration (30 s to 90 s) in the right plot.
Appendix 4.4. Three-dimensional stem plot from which required session duration can be determined when acceptable accuracy limits and practical MTS interval duration is selected for behaviours of 20% overall duration. Behaviours with short bout duration (1 s to 10 s) are represented in the left plot, and behaviours with medium bout duration (30 s to 90 s) in the right plot.
Appendix 4.5. Three-dimensional stem plot from which required session duration can be determined when acceptable accuracy limits and practical MTS interval duration is selected for behaviours of 75% overall duration. Behaviours with short bout duration (1 s to 10 s) are represented in the left plot, and behaviours with medium bout duration (30 s to 90 s) in the right plot.
Appendix 4.6. Three-dimensional stem plot from which required session duration can be determined when acceptable accuracy limits and practical MTS interval duration is selected for behaviours of 90% overall duration. Behaviours with medium bout duration (30 s to 90 s) are represented in the left plot, and behaviours with long bout duration (120 s to 300 s) in the right plot.