Int. J. Human-Computer Studies 79 (2015) 118–125

Contents lists available at ScienceDirect

Int. J. Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Studying interruptions and multitasking in situ: The untapped potential of quantitative observational studies

Scott R. Walter[a,*], William T.M. Dunsmuir[b], Johanna I. Westbrook[a]

[a] Centre for Health Systems and Safety Research, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia
[b] Department of Statistics, School of Mathematics and Statistics, University of New South Wales, Sydney, Australia

ARTICLE INFO

Article history:
Received 1 March 2014
Received in revised form
6 January 2015
Accepted 23 January 2015
Communicated by: E. Motta
Available online 4 February 2015

Keywords:
Interruption
Multitasking
Observational studies

ABSTRACT

Much of the large and growing body of literature on interruption and multitasking is motivated, in part, by a desire to reduce their negative effects in occupational settings, particularly those that are safety critical. Much of the existing knowledge has come from experimental studies, however, these do not necessarily generalize to non-experimental contexts. By virtue of being in situ, the results of observational studies are more generalizable, but internal validity remains an issue. Since many of the quantitative observational studies of interruption or multitasking to date have been largely descriptive, their full potential to contribute knowledge that informs practical improvements has been underutilized. We discuss ways to address threats to internal validity in quantitative observational studies through appropriate analysis with particular reference to workflow time studies, a form of direct observation. We also discuss the potential for more sophisticated analysis methods to both address some of the threats to internal validity and to provide more nuanced insights into the role and impacts of interruption and multitasking. In this way observational studies can contribute unique evidence to facilitate practical improvements to work practices and systems.

1. Introduction

A key motivation to understand interruptions and multitasking is to improve the accuracy and efficiency of work in occupational contexts. This is particularly true in safety critical settings such as air traffic control, aviation, healthcare, industrial process monitoring, and driving where error and inefficiency can have injurious or costly repercussions. In-depth knowledge of the role and impacts of interruptions and multitasking can inform improvements to workplace safety, practices and systems. Due to the complexity and heterogeneity of workflow and individuals in such settings, studying aspects of human work processes, such as interruptions or multitasking, present many challenges for quantitative study design and analysis.

Several approaches can be employed to study work processes including controlled experiments, computer simulation studies, and observational studies. Both experiments and simulations can be designed to control known and unknown sources of bias and thus achieve a high level of internal validity. However, the generalizability of results is limited by their similarity to non-experimental occupational settings, that is, they can lack sufficient external validity (Shadish et al., 2002). Some experimental studies have attempted to replicate interruptions or multitasking in contexts of interest, such as an office environment (Mark et al., 2008), cockpit (Latorella, 1999), motor vehicle (Watson and Strayer, 2010) or operating room (Liu et al., 2009); however, this becomes increasingly difficult for more complex and unpredictable settings such as hospital emergency departments (ED). Computer simulation studies provide a means to model interruptions or multitasking in more complex scenarios in a controlled way [see for example: (Lebiere et al., 2001; Sierhuis et al., 2007)], but this approach is limited by the accuracy of the necessary assumptions and, as with experiments, it can also be difficult to capture all the complexities of an uncontrolled setting. To date simulation studies of work in complex settings like EDs have focused on aspects such as patient flow and staffing, but not on interruptions or multitasking – an exception being a study (Gunal and Pidd, 2006) that simulated the effect of multitasking, in the sense of concurrent patient management, on departmental performance.

There are many types of observational studies that can be applied to investigate interruption and multitasking. Qualitative observational studies can provide insights about relationships, social dynamics and individual motivations and thought processes in a way that quantitative studies cannot, and this can be valuable when studying complex socio-technical settings. Nugus and Braithwaite (2010) used an ethnographic approach in an ED to understand the seemingly opposing

* Correspondence to: Centre for Health Systems and Safety Research, Australian Institute of Health Innovation, Faculty of Medicine and Health Sciences, Level 6, 75 Talavera Road, Macquarie University, New South Wales 2109, Australia. Tel.: +61 2 9850 2402; fax: +61 2 8088 6234.
E-mail address: scott.walter@mq.edu.au (S.R. Walter).

http://dx.doi.org/10.1016/j.ijhcs.2015.01.008
1071-5819/

factors of quality and organizational efficiency: a question which encompasses issues around multitasking and interruptions. Colligan and Bass (2012) used a combination of semi-structured interviews and direct observation to examine strategies that nurses used to handle interruptions.

While all types of study can contribute important knowledge about interruption and multitasking, in this article we focus on quantitative observational studies for several reasons. They can be conducted in the setting of interest, hence making their results generalizable to at least that context or others that are similar (Black, 1996). For example, a study of medication administration errors found that the risk and severity of error increased with the number of times the administration was interrupted (Westbrook et al., 2010b). Observing interruptions of nurses in situ provides a more accurate assessment of their potential impact on nurses' work than results from experiments or simulations. There may also be ethical constraints on conducting experiments or interventions in safety critical settings where the effect of unintended negative consequences could be serious. The same restriction is less of an issue for observational studies where the data collection process aims to have minimal impact on the context under study. However, a major drawback to the quantitative observational approach is that it can be difficult to establish internal validity and to date this has proven restrictive to the rate of knowledge generation about interruption and multitasking, particularly in healthcare.

The majority of quantitative observational studies of interruptions or multitasking are situated in medical contexts and, as noted previously (Coiera, 2012; Grundgeiger and Sanderson, 2009), most of these have essentially taken a “counting” approach by simply summarizing counts, rates and proportions. A select few healthcare studies have taken a more advanced appraoch. The previously mentioned medication administration study used a multivariate analysis to find an association between interruption and error (Westbrook et al., 2010b), while another study of intensive care unit staff used eye tracker technology and a multilevel multivariate model to analyse resumption lag following an interruption (Grundgeiger et al., 2010).

While the quantitative observational approach is well suited to healthcare, it is also applicable in other domains. Several studies of information workers have used this approach to examine concurrent task management (Czerwinski et al., 2004; Gonzalez and Mark, 2004), and Loukopoulos et al. (2001) conducted a study of interruption and task interleaving among pilots by observing their activities from the cockpit jumpseat. In an observational study of drivers, Strayer and Drews (2006) assessed the association between concurrent hand held cell phone use while driving and failure to stop at an intersection.

The need to advance the research agenda for interruptions and multitasking in healthcare has been recently noted (Westbrook, in press), and there is clearly considerable scope for more rigorous observational studies to contribute practically useful knowledge to occupational domains, whether healthcare or otherwise. In this paper we aim to expound the ways in which the design, data collection and analysis of quantitative observational studies of interruption and multitasking can be improved from current practice. In particular we discuss fundamental issues with the internal and external validity of observational research in reference to interruption and multitasking, and the ways in which these issues can be mitigated through the application of existing statistical techniques. We also point out areas in which new statistical developments are needed and outline ways forward for each. Where possible, we illustrate these points via a hypothetical case study.

2. Workflow time studies

There are many approaches that can be employed to record an individual's work process, as discussed at length by Lopetegui et al. (2014). The workflow time study approach (Lopetegui et al., 2014) is a type of time and motion study that offers many advantages over other non-experimental methods applicable to work processes. It involves an external observer shadowing a participant and recording time-stamped information about their tasks and interactions to create a continuous record of the work process. It has its roots in Mintzberg's structured observation method (1970) and is also similar to systematic direct observation used in timed-event sequential analysis in psychology (Bakeman and Gottman, 1997; Chorney et al., 2010) in that it involves recording behaviour in an uncontrolled setting according to predefined operational definitions. The additional emphasis in workflow time studies is on capturing a continuous record of behaviour. It is distinct from an ethnographic approach where observed interaction or behaviour is categorized during the analysis phase (Atkinson and Hammersley, 1994). Workflow time studies have been applied to interruption and multitasking in the domains of healthcare (Weigl et al., 2011; Westbrook et al., 2010a), aviation (Loukopoulos et al., 2001) and human–computer interaction (Gonzalez and Mark, 2004; Mark et al., 2012; Su and Mark, 2008).

The continuous recording of data increases the potential to capture work complexity compared to work sampling or self-report approaches such as diary studies (Mintzberg, 1970). It is also less prone to bias than work sampling (Finkler et al., 1993) or self-report. While audio or video recording can provide an accurate continuous record of a work process, these can easily capture non-participants and the need to seek consent from all those recorded can be prohibitive. In addition, workflow time studies open up the analysis possibilities to a wide range of existing techniques, each of which has the potential to provide innovative insights. Hence we focus on this observational approach and the ways in which it can minimize threats to internal validity and can broaden the scope for statistical analyses applicable to observational data on interruptions and multitasking.

3. Internal validity

One of the main challenges in quantitative observational studies is to generate internally valid results, that is, results that are not biased. This is particularly so in complex settings where there is a network of intertwined factors at play and separating out the influences of particular factors requires addressing the many threats to internal validity. In this section we outline some of those threats and how they can be mitigated with reference to workflow time studies.

3.1. Defining interruptions and multitasking

There is much heterogeneity in the definitions of interruptions and multitasking. Many studies provide no explicit definition, while others attempt to bring some precision to particular terms, such as Trafton et al. (2003) often cited ‘anatomy of an interruption’ (Fig. 1). The study of interruption and multitasking is now beset with inconsistency, with some terms having been defined to have several different meanings, and some concepts described by several different terms. For example, with reference to Trafton

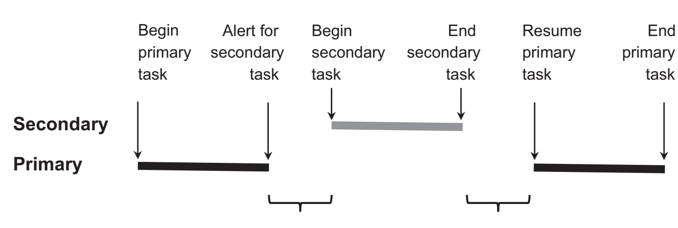

Fig. 1. Trafton et al.'s anatomy of an interruption.

et al.'s model, an *interruption* has been defined as the alert for the secondary task (Chisholm et al., 2001; Czerwinski et al., 2004; Mache et al., 2012), the secondary task itself (Li et al., 2012), or the whole sequence depicted in Fig. 1 (Boehm-Davis and Remington, 2009; Weigl et al., 2011). *Multitasking* has also been defined in several different ways. The notions of *concurrent* multitasking (also called *dual task performance*), *interleaved* multitasking (also called *task-switching*) and *sequential* multitasking (Adler and Benbunan-Fich, 2012; Loukopoulos et al., 2009) have been combined under a unifying theory in which these concepts represent different places on a continuum depending on the rate of switching between tasks (Salvucci et al., 2009). Yet to complicate things, task-switching that is externally triggered is sometimes called multitasking in the experimental literature [e.g. (Katidioti and Taatgen, 2014)] while generally called interruption in the healthcare literature. While several papers have discussed the panoply of terms and definitions (Brixey et al., 2007; McFarlane, 1997; Sasangohar et al., 2012), there is often an assumption that a single definition is both possible and desirable [see for example (McFarlane, 1997)].

As a way through the semantic imbroglio there are several considerations for future observational studies. First, definitions should be developed specific to the context and the research hypotheses. In complex non-experimental settings people juggle competing demands in a wide range of creative ways such that behaviours that might be considered 'interruption', 'task-switching' or 'multitasking' according to previous definitions, can occur in unlikely sequences or may be mixed together in ambiguous ways. While there are many proponents of a universal definition of interruption (Brixey et al., 2007; Grundgeiger and Sanderson, 2009; Sasangohar et al., 2012) or multitasking (Salvucci et al., 2009), if such definitions were possible, every time the definition was applied in a new context it would necessarily have to be reinterpreted and re-operationalized, hence defeating the purpose of a universal definition. Secondly, definitions should be as precise as possible to minimize measurement error or observer bias. In a fast moving fluid environment it is essential to have operational definitions that distinguish what is and is not considered an interruption or a multitask as clearly as possible. This is necessary so that observers can translate observed behaviour into a record of the work process in a repeatable way (Hintze et al., 2002). This also supports transparency and comparability when publishing results. Developing and operationalizing definitions can be an iterative process by which definitions are tested (through piloting), adjusted and retested until they can be applied in a way that minimizes bias and error. A final and optional consideration is that definitions can be chosen according to an underlying construct of interest (Grundgeiger and Sanderson, 2009). We now introduce a hypothetical case study to illustrate these points, and this will be used throughout the paper to elucidate subsequent points. For the purpose of illustration, the study is somewhat simplified. A study aims to determine factors associated with non-resumption of interrupted tasks among doctors in an emergency department (ED). Using a similar idea to multiple resource theory (Wickens, 2002) the investigators hypothesize that non-resumption is a failure of prospective memory and occurs when the context is demanding enough that insufficient cognitive resources remain to recall the intention to resume a task. ED doctors often work in open departments as part of a team and there are many events that could potentially be included in the definition of an interruption. For a task to be at risk of not being resumed, it first has to be suspended prior to completion. This task-switching may be externally or internally triggered, but the researchers decide that knowing about external triggers is more informative for improving practice. Hence the study team decides to define an interruption as a switch from one task to another prior to completion of the original task, and where the switch is triggered by an external event. An external event is defined to include anything specifically directed towards the doctor including phone calls, questions, computer alerts and pager calls, but excluding equipment alarms, nearby conversations, other people's phones ringing and so on.

Having developed study-specific definitions and operationalized them, the next consideration is to assess the extent to which definitions are reliably applied.

3.2. Intra- and inter-rater reliability

When data collection relies on an observer interpreting what they see and hear, there is potential for variation in how the definitions are applied from one observation period to the next for the same observer (*intra*-rater reliability), and from one observer to the next (*inter*-rater or *inter*-observer reliability). In addition to the definitional precision discussed above, reliability is usually optimized through observer training and a quantitative measure of agreement. For inter-rater reliability this simply means having two (or more) observers record data while observing the same scenarios. Intra-rater reliability is less easy to test as it ideally involves an observer of the same scenarios at different points in time.

Establishing a sufficient level of reliability requires a means to quantify it. The original inter-rater methods that arose from the field of psychology assess agreement between two raters classifying the same entity into a set number of categories [e.g. (Cohen, 1960)]. Although some of the univariate measures, such as Cohen's kappa, are often applied in observational studies of work processes (Lopetegui et al., 2013), these methods have two main limitations. The first is that they do not take into account the temporal ordering of tasks: the time stamp of each task and its place in the sequence of tasks are important considerations for reliability. Secondly they cannot assess agreement simultaneously between multiple variables. For example, a task may have several characteristics such as start time, duration, type (e.g., documentation) and interruption status and ideal agreement requires all characteristics to match. Determining whether data from two observers agrees on all of these attributes together is beyond the scope of existing methods. High kappa values do not necessarily mean there is good intra- or inter-rater reliability for all aspects of the collected data. The TimeCaT software includes a multidimensional measure of both inter- and intra-rater reliability by separately comparing total task time, total task count, click accuracy and sequence similarity (TimeCaT 3.9, 2013). Many methods exist for determining similarity between two strings of data or, equivalently, two multivariate records, including probabilistic record linkage (Herzog et al., 2007). Such methods have the potential to be adapted to quantify agreement in a way that takes into account the number of variables being simultaneously compared – perfect agreement becomes increasingly difficult when more variables have to match. In a similar way to Cohen's kappa, these approaches can also indicate the extent to which the level of agreement exceeds that expected due to chance alone. For example, probabilistic linkage could be used to identify the best matching unique record pairs and the sum of total match weights for all the 'best' pair matches aggregated into an overall score. To compare the likelihood of that score occurring by chance, a p-value for this total could then be obtained via a Monte Carlo permutation approach, that is, by recalculating the score for random shuffles of the data to generate a sampling distribution to which the original score can be compared.

These techniques can facilitate minimization of bias or error during direct observations which can then enable more accurate analysis results.

3.3. The importance of capturing covariates in uncontrolled settings

In an uncontrolled setting there may be many factors that simultaneously yet differentially influence a particular outcome of interest. In an experiment these factors are controlled to isolate a particular effect. Otherwise, where these factors are quantifiable, they can be analysed concurrently to separate out the effects of each. There are many well established techniques for doing this, multivariate regression being one of the most well known. In this context these concurrent factors are often referred to as covariates. The workflow time study approach enables the simultaneous collection of data on many covariates. In this section we consider the importance of covariates in terms of two broad analysis approaches.

The first is the hypothesis driven approach which aims to test the effect of one or more predetermined factors on an outcome. In our case study example, as described in Section 3.1, this might be the effect of an interrupting task involving patient resuscitation (a binary variable) on the risk of non-resumption of the original task (also binary). This is simplified to illustrate that a demanding interrupting task may be hypothesized to be more likely to cause non-resumption of the original task. Due to the non-randomized nature of the data, the relationship between these variables may be confounded by other factors such as the experience level of the doctor. A confounder is a variable that is separately related to exposure and outcome, and failure to account for it can result in bias in the effect of interest (Greenland et al., 2008). Of the many ways to deal with confounding, the most applicable to quantitative observational data is multivariate modelling where the effect of interest and all potential confounders are included as covariates. Methods for selecting covariates for regression adjustment are well covered elsewhere (Greenland, 1989; Schisterman et al., 2009). It is possible to try to collect information on as many confounders as possible, then apply variable selection techniques to find the most relevant factors. Alternatively, confounders may be hypothesized a priori and only information on those predetermined factors collected and adjusted for. Weigl et al. (2012) provide an example of this kind of analysis where they examined the association between interruptions and perceived workload among hospital doctors, while adjusting for time of day and doctor seniority as confounders.

Alternatively, an exploratory approach aims to identify factors associated with the outcome of interest as opposed to testing a particular hypothesis. The significant variables are distilled down from the set of all available variables via a model building process that aims to find the model that best explains the data [see for example, Hosmer and Lemeshow's purposeful selection of covariates, a model building process that aims to improve on automated stepwise methods (Hosmer and Lemeshow, 1999), Section 5.2]. This type of analysis is relevant to complex socio-technical settings where there are many potential factors and little may be known about their interrelationships. In the case study this would apply to the question: what factors are associated with non-resumption of an interrupted task? In addition to the variables mentioned for the hypothesis driven approach, the researchers also consider characteristics of the interrupting task (type, interrupting person, duration, arrival time during primary task). Other covariates may further be constructed from the data such as the interruption rate in a time window preceding the interrupted task, or the number of non-resumed tasks accumulated to a given point in time. Grundgeiger et al. (2010) and Walter et al. (2014) each present examples of a model building approach.

3.4. Maximizing internal validity through analysis

Regression modelling, as mentioned in Section 3.3, is a flexible way of analysing interruptions or multitasking situated in work processes. The possibilities are manifold, but there are certain aspects of regression that are important for minimizing bias and have received limited attention in observational studies of interruption and multitasking to date. The importance of covariates adjustment has already been discussed, so we now also consider autocorrelation, clustering and unmeasured confounding. In data where the outcome is temporally ordered there is often correlation between the value of an outcome variable at a given time and previous values of that variable, known as autocorrelation. This is commonly dealt with by including autoregressive error terms or by including lagged values of the outcome as covariates in the regression model. Failure to account for autocorrelation can have serious impacts on accuracy and precision (Pollitt et al., 2012). As a simple example, in a study of factors associated with a clinician's choice to switch tasks or concurrently multitask when triggered by an external event, Walter et al. (2014) included the choice at the previous trigger as a potential covariate.

Another form of correlation is due to clustering. This occurs when outcome values show correlation within certain subgroups or clusters and represents another potential source of bias if ignored, particularly for standard error estimation (Diggle et al., 2002). There are a number of ways to account for this in a regression context with generalized estimating equations and random effect models being two common approaches. In general, a particular work unit or group will have its own practices and team culture, and individuals within each setting will have their own ways of working. Hence individuals and groups are two potential levels of clustering that may need to be addressed in a task-level analysis, although the levels of clustering will be specific to the study design and setting. Grundgeiger et al. (2010) and Walter et al. (2014) present examples where random intercepts models were used to address clustering within individuals.

Multivariate modelling is an effective way to adjust for the effects of known confounders that involves optimizing precision and confounder-related bias while avoiding over adjustment and unnecessary adjustment. Techniques also exist for taking into account the effect of unmeasured confounders (Hougaard, 1995; Lin et al., 1998). While less commonly performed, these methods provide a way to increase evidence for (or against) a causal relationship. If an estimated effect is relatively resilient to a range of assumptions about unmeasured confounding then this strengthens the evidence of a real relationship not due to other factors (Lin et al., 1998).

4. External validity

External validity is the extent to which study results are relevant to settings other than the original study setting. This is important in that the point of most studies is to generate knowledge that is generalizable. We outline two considerations in this vein: the influence of external observers on the participants and ensuring there is sufficient statistical power to detect genuine effects of interest.

4.1. Reactivity

The presence of an observer has the potential to influence the way an observed person behaves. For instance, being observed may promote productivity or better adherence to official procedures. This is often referred to as the Hawthorne effect and the presence of such an effect can introduce bias. The existence of this phenomenon in the original study of the Hawthorne Works in Chicago has been subsequently questioned or contradicted, the main issue being the lack of adjustment for other factors influencing workers' productivity. Jones (1992) performed a multivariate reanalysis of the original data from the relay assembly test room and reported no evidence of a Hawthorne effect after adjusting for other possible confounding factors. A reanalysis of the illumination experiments found no evidence of an immediate response to changes in light conditions (Levitt and List, 2011). However, there is still potential for observers to have an influence in other settings. While the observer ideally aims to watch a participant from a fly-on-the-wall perspective, it is unavoidable that they themselves become a part of the setting.

Simple tactics to mitigate the potential for observer influence are to avoid interaction between observer and subject and for the observer to remain far enough away so as not to encroach on the subjects' performance of their work (Weigl et al., 2011). Also, assurance that performance is not being assessed and that any errors will be recorded in a non-identifiable way will help to minimize any sense of being under scrutiny. A period of acclimatization, prior to beginning observations proper, can allow subjects to become somewhat comfortable with being directly observed. Conducting observation sessions throughout an extended period, several months say, may also reduce to possibility of sustained behaviour changes (Westbrook and Ampt, 2009).

4.2. Sufficient power

A precursor to having generalizable results is to have sufficient power to detect real effects or estimate the prevalence of events – such as interruptions – with sufficient precision. This is traditionally achieved through sample size calculation prior to commencement of observations. Most standard sample size formulae originate from the domain of experiments. Less straight forward methods exist that are applicable to non-experimental studies; however, in general the more the factors being collected and analysed, the more challenging the sample size estimation becomes. This can be compounded in the absence of any prior knowledge or evidence to form the basis of the calculation assumptions.

In observational studies of work processes, sample size calculation is rarely mentioned, yet for the resource intensive workflow time study method this could be worthwhile. Applying a more-is-better approach may not necessarily result in increased power since an expanded sample may capture a more diverse group of subjects, or if the observations are carried out over an extended period of time temporal variation may be introduced. Capturing and accounting for these additional sources of variability can increase generalizability, but can also considerably augment the sample size required to maintain a given level of precision due to the inverse relationship between the number of parameters estimated by a regression model and the precision of the estimates. Also the context specific nature of observational studies means that there is a limit to how widely their results can be generalized.

Sample size determination methods exist for multivariate models, multi-level models and time series; however, for some types of analyses no directly applicable methods have been developed. In place of developing new methods it is possible to adapt existing methods by making some simplifying assumptions, or to apply several methods and use the most conservative estimate. Even if it is not possible to generate a precise sample size estimate it can still be worthwhile ensuring that there is sufficient power to assess the main hypotheses of interest. Previous studies can be informative in determining sample size, but in the absence of prior relevant information another possibility is to carry out a period of observations or collect some pilot data, perform interim analyses and then use those results to update the sample size calculation based on the pilot effect sizes.

We illustrate how a researcher might go about determining sample size for a workflow time study by revisiting the case study. The researchers are interested in the relationship between a task not being resumed and the interruption of the task being caused by a resuscitation call, both binary variables. The analysis plan is to use logistic regression on all interrupted tasks with non-resumption as the outcome and resuscitation status of the interruption as the main covariate of interest. It is expected that the type of task and the time since task beginning to interruption will have confounding effects and will be included as additional covariates. From previous data the non-resumption rate has been estimated at 5% of interrupted tasks, interruptions affect 20% of tasks, and doctors complete 12 tasks per hour on average. The researcher starts with a simple approach that ignores the confounders and generates a sample size estimate of 1647 tasks, which can be translated to about 168 h. There is no prior information on the distribution of possible covariates or on the direction or size of their effects, so the researchers recalculate including the two confounders, but with a range of plausible distributions and effect sizes for each. This gives estimates from 103 to 225 h. At most the study has resources for 200 h of observation so it is decided to refine the multivariate calculation after obtaining estimates of confounder covariate effects from the initial 20 h of observation. The refined calculation gives an estimate of 145 h. Since this is lower than the initial estimate of 168, the researchers decide to take a conservative approach to ensure sufficient power, but to conserve at least some resources by collecting 170 h of observations.

5. Scope for analytic innovation in quantitative observational studies

Workflow time studies enable a wide range of analysis methods relevant to observational studies of interruption and multitasking but such methods have been under used or not used thus far. We describe the application of a selection of these techniques in this section. These are well described elsewhere and we only provide a brief outline of each.

5.1. Linking interruption and multitasking to outcomes: association and causation

The aim of much of the research on work processes is to examine associations between phenomena such as interruptions or multitasking and particular outcomes such as error or inefficiency. Due to the dynamic and complex nature of the settings in which interruptions and multitasking are endemic, establishing a link between a particular interruption, say, and a particular error is analytically challenging. While ideally we would like to know the precise cognitive steps that led to each error, i.e. the causal explanation, for the most part a quantitative observational approach identifies only observable factors that are potentially causative of errors, i.e. the causal description (Shadish et al., 2002).

There is much literature on what constitutes evidence for causality, including a number of suggested minimum conditions [e.g. Hill, 1965]. Three widely used conditions originating with John Stuart Mill (Cook and Campbell, 1979) are that the cause should precede the effect, the purported cause and effect are related, and other possible explanations for the relationship can be eliminated. Although causality cannot be established with any certainty, and there is no neat road map for doing so, observational studies that attempt to establish these three conditions may identify at least some of the component causes of error (Rothman et al., 2008). The point of statistical regression models is to quantify associations, hence effects identified through modelling satisfy the second condition, while adjustment for confounders helps to rule out other explanations for an observed association (third condition). Addressing the first condition requires incorporation of some measure of temporal ordering into the analysis.

A relatively simple approach to linking interruption and error, used in several studies to date, is to examine the association between rates aggregated over some time period. For example, Flynn et al. (1999) found an association between the rate of interruptions per half hour and the rate of medication dispensing errors aggregated over the same unit of time. This could similarly be applied to some aggregated measure of multitasking. While appealingly simple, we do not know whether the interruptions had anything to do with the errors, only that their occurrence rates were somewhat correlated. Many tasks may have been carried out during each half hour period, yet we have no sense that interruptions were temporally close to and preceding errors or whether there was another factor driving both. For example increased workload may amplify the frequency of interruptions and perceived excessive workload may increase the propensity for error. In terms of system design there is not enough information to indicate how to make changes to reduce errors.

Westbrook et al. (2010b) used a more targeted approach where the association between interruptions and errors was examined only during one particular type of task: medication administration. This is similar to a single duration measurement approach (Lopetegui et al., 2014). While this more clearly establishes the temporal proximity (although not ordering) of interruptions and errors, it ignores much of the work process that could also have contributed to the occurrence of both.

With the benefit of having a continuous record of the work process it is possible to explicitly examine temporal ordering of interruption (or multitasking) and error. This may be done by including covariates that capture information about interruptions or multitasking that precede each error. An example of this might be the interruption rate in a local time window preceding each error, or the time since last interruption. Alternatively, past or future values of a covariate may be included to examine associations at different lead or lag intervals. The effect of an interruption on reaction time has been shown to persist for a period after the interruption (Altmann and Trafton, 2007) and this is an obvious scenario in which lagged relationships could be modelled. Schildcrout and Heagerty (2005) discuss this for binary outcomes, and there is much literature on the related approach of distributed lag modelling (Almon, 1965). Since these models can also establish strength of association and adjust for other possible explanatory variables, they are a means by which a causal description can be established.

5.2. Bridging the gap between experimental psychology and observational studies

Observational studies in uncontrolled settings cannot isolate cognitive phenomena related to interruption or multitasking in the same way a controlled experiment can. One of the few studies to attempt to do so used eye tracking technology to directly measure resumption lag following interruptions (Grundgeiger et al., 2010), something only previously measured in experiments. However, in this section we broadly outline a set of modelling techniques by which inference about underlying psychological constructs (latent variables) may be drawn from directly observed behaviour. To illustrate via the case study, abandoning an interrupted task may be a failure of prospective memory and the researchers hypothesize that this is due to excess cognitive resources being consumed by the interrupting task. Hence whether the interrupted task is abandoned is considered a binary realization of the unobserved, or latent, level of available cognitive resources at that time. Similarly, a participant's choice between interruption and multitasking when triggered is hypothesized to represent the same construct, with a choice of interruption indicating less available resources than if multitasking was chosen. Other observable aspects of the work process can also be defined as realizations of the same construct to create a range of observed measures of that construct. The investigators are interested in how workload relates to the level of cognitive resources and so collect several workload measures: individual heart rate variability, department level patient load, and number of patients concurrently managed by each participant. Within the broad framework of structural equation modelling it is possible to assess the relationship between available cognitive resources and workload by modelling the observed realizations of these latent factors. As a further example of how observed behaviour can be used to make inference about latent constructs, a study of primary school children used observations of 20 different behaviours related to memory deficits and 12 validated tests of working memory to draw conclusions about the relationship between two underlying constructs: cognitive working memory and behavioural working memory (Alloway et al., 2009). The estimated correlation between the two factors was reported as 0.52. This broad approach applied to observational data of work processes may form an important means by which to study relationships among psychological phenomena related to interruption and multitasking in non-experimental settings.

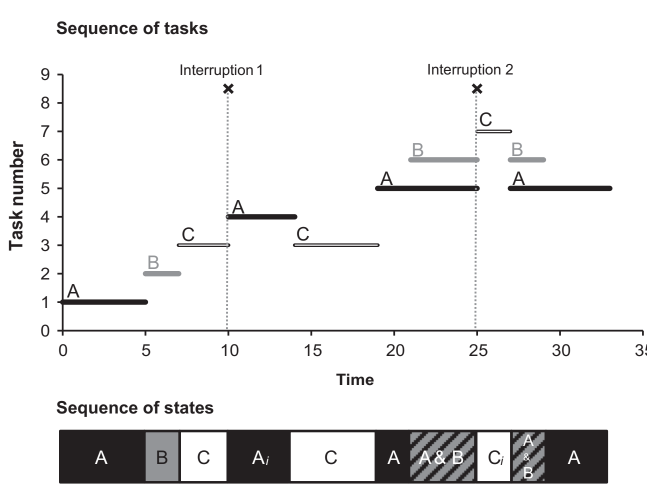

Fig. 2. A sequence of three different task types (A, B or C) involving interruption and multitasking with an example of how the process can be conceptualized as a sequence of states.

5.3. Other analysis possibilities

The continuous record of a work process can be conceptualized as a series of states. For example, conversing may be considered one particular state, documenting could be another. Fig. 2 provides an example illustration of how a sequence of tasks may be conceptualized as a sequence of states. This opens the way for the use of Markov models. A relatively simple possibility is then to model the probability of transitioning to various states given the current state and, optionally, given a certain number of previous states. Definitions of states can then be defined to be relevant to interruption and multitasking. For example, the probabilities of transitioning from conversing to some other task could be compared where the conversation is interrupted versus when it is not, thus capturing the effect of interruptions on workflow. There are many Markov models that may be relevant to particular hypotheses, and many relevant applications may be found in the related field of sequential behaviour analysis (Bakeman and Gottman, 1997). In one of the few examples applied to interruption and multitasking, Su and Mark (2008) use Markov transition probabilities to examine task switching (i.e., interleaved multitasking) in sequences of tasks grouped into communication chains.

A final analysis consideration related to the impact of particular events on task length is the phenomenon of length bias. If events such as interruptions occur at random points in time, the likelihood of a task being interrupted is proportional to its duration. Thus it is not valid to compare lengths of interrupted and uninterrupted tasks to assess whether interruptions have an effect on task length. Instead, it is necessary to estimate the length bias adjusted expected task length for a given number of interruptions, assuming there is no interruption effect, and compare this to the observed lengths. A significant difference between observed and expected values provides evidence for an interruption effect. One type of interruption effect is to lengthen the time taken to complete a task through resumption lag. This has been studied through direct timing of resumption lag in experimental studies (Altmann and Trafton, 2004) and has been assessed in an occupational setting (Grundgeiger et al., 2010); however, analytic methods to assess an interruption effect can also incorporate other types of effects such as task shortening. A method to do this was proposed by Brown and Dunsmuir (2010) as part of Westbrook et al.'s, 2010a study of emergency department doctors, although there is considerable scope for extension.

6. Discussion

We have outlined the considerable scope for quantitative observational studies to contribute important evidence about interruption and multitasking specific to occupational settings through the use of workflow time studies and increased analytic rigour and innovation. Arguably, the main motivation for studying interruption and multitasking in such settings is to improve performance by reducing error and inefficiency. The role of quantitative observational studies is manifold in this respect. They can be used to gain an understanding of the way interruption and multitasking function in a particular setting and to then inform the nature of improvements to practice. Many of the observational healthcare studies of interruption and multitasking fall into this category and such studies are necessary when little is known about a complex setting. This is analogous to the exploratory approach described in Section 3.3. Observational studies can also be used to test particular hypotheses generated by previous research and related to a proposed set of changes. Further, observational studies can be applied to assess implemented changes or interventions in situ, that is, quasi-experimental observational studies [e.g. (Weigl et al., 2014)]. The latter two applications are more akin to the hypothesis driven approach previously described. While the risk of unintended negative effects of a poorly informed wide-spread intervention is potentially disastrous in safety critical settings, elsewhere it may be more possible to intervene in work practice at a small scale and see what happens. For example Mark et al. (2012) cut off email to a group of scientific researchers and assessed the impact on their use of interleaved multitasking. Regardless of the way observational studies are applied, the points outlined in this paper apply equally.

Where experiments can isolate specific aspects of interruption and multitasking, observational studies situated in occupational contexts can take a broader perspective. Interruptions may be a contributing factor to some negative outcomes, yet in terms of system design it may be more useful to identify what drives interruptions in the first place. If interruptions are symptomatic of high workload then a focus on managing workload may be more beneficial than focusing on interruptions alone. Improvement to a complex system may also require broadening from observations of individuals to observing the whole system, as discussed by Harr and Kaptelinin (2007). An environment characterized by interruptive communication may seem suboptimal for the individual, but could be the most efficient means of timely information transfer to ensure successful operation of the team. Conversely, reducing the level of interruptions at an individual level may not result in system wide improvement.

We have outlined many analysis techniques including models that provide insight about unobserved variables and causal relationships. However, there is a limit to what can be learned from quantitative studies of the type we have discussed. Observers can only capture a certain amount of quantitative information; hence there is a role for qualitative studies in capturing more nuanced details of interactions, as exemplified in the qualitative studies described in the introduction (Colligan and Bass, 2012; Nugus and Braithwaite, 2010). A further limitation of workflow time studies is that they can be resource intensive, with previous such studies often requiring several hundred hours of observation to capture sufficient errors and covariate information (Westbrook et al., 2010a, 2010b). This needs to be weighed against the proposed benefit of well-informed changes and the cost of poorly informed changes.

The real strength of well conducted observational research is that it can be situated in working contexts. This has the potential to provide knowledge of genuine use for improving practice, particularly in settings where the negative effects of interruption and multitasking could be costly. The many possibilities outlined in this paper underscore the untapped potential of this type of research in the study of interruption and multitasking as well as work processes in general.

Acknowledgements

This work was funded by ARC Discovery Grant DP110100090 and NHMRC Program Grant 1054146. SRW is funded by a NHMRC Postgraduate Scholarship. The authors would like to thank the editor, the anonymous reviewers, Dr. Heather Douglas and Dr. Arathi Sriprakash for their valuable feedback.

References

Adler, R.F., Benbunan-Fich, R., 2012. Juggling on a high wire: multitasking effects on performance. Int. J. Hum. Comput. Stud. 70 (2), 156–168.
Alloway, T.P., Gathercole, S.E., Kirkwood, H., Elliott, J., 2009. The working memory rating scale: a classroom-based behavioral assessment of working memory. Learn. Individ. Differ. 19 (2), 242–245.
Almon, S., 1965. The distributed lag between capital appropriations and expenditures. Econometrica 33 (1), 178–196.
Altmann, E.M., Trafton, J.G., 2004. Task interruption: resumption lag and the role of cues. In: 26th Annual Meeting of the Cognitive Science Society. Chicago.
Altmann, E.M., Trafton, J.G., 2007. Timecourse of recovery from task interruption: data and a model. Psychon. Bull. Rev. 14 (6), 1079–1084.
Atkinson, P., Hammersley, M., 1994. Ethnography and participant observation. In: Denzin, N.K., Lincoln, Y.S. (Eds.), Handbook of Qualitative Research. Sage Publications, Inc., Thousand Oaks, CA, USA, pp. 248–261.
Bakeman, R., Gottman, J.M., 1997. Observing Interaction: An Introduction to Sequential Analysis. Cambridge University Press, Cambridge, UK.
Black, N., 1996. Why we need observational studies to evaluate the effectiveness of health care. BMJ 312 (7040), 1215–1218.
Boehm-Davis, D.A., Remington, R., 2009. Reducing the disruptive effects of interruption: a cognitive framework for analysing the costs and benefits of intervention strategies. Accid. Anal. Prev. 41 (5), 1124–1129.
Brixey, J.J., Robinson, D.J., Johnson, C.W., Johnson, T.R., Turley, J.P., Zhang, J., 2007. A concept analysis of the phenomenon interruption. Adv. Nurs. Sci. 30 (1), E26.
Brown, B.M., Dunsmuir, W.T.M., 2010. Adjusting for length biased sampling when assessing the impact of interruptions on total time on task in a hospital emergency department setting. Qual. Saf. Health Care 19 (4), 284–289 (web only supplement).
Chisholm, C.D., Dornfeld, A.M., Nelson, D.R., Cordell, W.H., 2001. Work interrupted: a comparison of workplace interruptions in emergency departments and primary care offices. Ann. Emerg. Med. 38 (2), 146–151.
Chorney, J.M., Garcia, A.M., Berlin, K.S., Bakeman, R., Kain, Z.N., 2010. Time-window sequential analysis: an introduction for pediatric psychologists. J. Pediatr. Psychol. 35 (10), 1061–1070.
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46.
Coiera, E., 2012. The science of interruption. BMJ Qual. Saf. 21 (5), 357–360.
Colligan, L., Bass, E.J., 2012. Interruption handling strategies during paediatric medication administration. BMJ Qual. Saf. 21 (11), 912–917.
Cook, T.D., Campbell, D.T., 1979. Quasi-Experimentation: Design & Analysis Issues for Field Settings. Houghton Mifflin, Boston, MA, USA.
Czerwinski, M., Horvitz, E., Wilhite, S., 2004. A diary study of task switching and interruptions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI 2004. ACM. Vienna, Austria, pp. 175–182.
Diggle, P.J., Heagerty, P., Liang, K.-Y., Zeger, S.L., 2002. Analysis of Logitudinal Data, 2nd ed. Oxford University Press, Oxford UK.
Finkler, S.A., Knickman, J.R., Hendrickson, G., Lipkin Jr, M., Thompson, W.G., 1993. A comparison of work-sampling and time-and-motion techniques for studies in health services research. Health Serv. Res. 28 (5), 577–597.
Flynn, E.A., Barker, K.N., Gibson, J.T., Pearson, R.E., Berger, B.A., Smith, L., 1999. Impact of interruptions and distractions on dispensing errors in an ambulatory care pharmacy. Am. J. Health Syst. Pharm. 56 (13), 1319–1325.
Gonzalez, V.M., Mark, G.J., 2004. "Constant, Constant, Multi-tasking Craziness": Managing Multiple Working Spheres, CHI2004. ACM, Vienna, Austria.
Greenland, S., 1989. Modeling and variable selection in epidemiologic analysis. Am. J. Public Health 79 (3), 340–349.

Greenland, S., Rothman, K.J., Lash, T.L., 2008. Measures of effect and measures of association. In: Rothman, K.J., Greenland, S., Lash, T.L. (Eds.), Modern Epidemiology. Lippincott Williams & Wilkins, Philadelphia, USA.
Grundgeiger, T., Sanderson, P., 2009. Interruptions in healthcare: theoretical views. Int. J. Med. Inf. 78 (5), 293–307.
Grundgeiger, T., Sanderson, P., MacDougall, H.G., Venkatesh, B., 2010. Interruption management in the intensive care unit: predicting resumption times and assessing distributed support. J. Exp. Psychol. Appl. 16 (4), 317–344.
Gunal, M.M., Pidd, M., 2006. Understanding accident and emergency department performance using simulation. In: Perrone, L.F., Wieland, F.P., Liu, J., Lawson, B.G., Nicol, D.M., Fujimoto, R.M. (Eds.), Winter Simulation Conference. Monterey, CA, USA.
Harr, R., Kaptelinin, V., 2007. Unpacking the Social Dimension of External Interruptions. ACM, Florida, USA, pp. 399–408 (GROUP '07).
Herzog, T.N., Scheuren, F.J., Winkler, W.E., 2007. Data Quality and Record Linkage Techniques. Springer, New York, USA.
Hill, A.B., 1965. The environment and disease: association or causation? In: Proceedings of the Royal Society of Medicine, pp. 295–300.
Hintze, J.M., Volpe, R.J., Shapiro, E.S., 2002. Best practices in the systematic observation of student behaviour. In: Thomas, A., Grimes, J. (Eds.), Best Practices in School Psychology. National Association of School Psychologists, Washington DC, USA, pp. 933–1006.
Hosmer, D.W., Lemeshow, S., 1999. Applied Survival Analysis. John Wiley & Sons, New York, USA.
Hougaard, P., 1995. Frailty models for survival data. Lifetime Data Anal. 1 (3), 255–273.
Jones, S.R.G., 1992. Was there a Hawthorne effect? Am. J. Sociol. 98 (3), 451–468.
Katidioti, I., Taatgen, N.A., 2014. Choice in multitasking: how delays in the primary task turn a rational into an irrational multitasker. Hum. Factors 56 (4), 728–736.
Latorella, K.A., 1999. Investigating Interruptions: Implications for Flightdeck Performance. National Aeronautics and Space Administration, Langley Research Center, Langley, USA.
Lebiere, C., Anderson, J.R., Bothell, D., 2001. Multi-tasking and cognitive workload in an ACT-R model of a simplified air traffic control task. In: Proceedings of the Tenth Conference on Computer Generated Forces and Behavioral Representation. Norfolk, VA, USA.
Levitt, S.D., List, J.A., 2011. Was there really a Hawthorne effect at the Hawthorne plant? An analysis of the original illumination experiments. Am. Econ. J.: Appl. Econ. 3 (1), 224–238.
Li, S.Y.W., Magrabi, F., Coiera, E., 2012. A systematic review of the psychological literature on interruption and its patient safety implications. J. Am. Med. Inf. Assoc. 19 (1), 6–12.
Lin, D.Y., Psaty, B.M., Kronmal, R.A., 1998. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics 54 (3), 948–963.
Liu, D., Grundgeiger, T., Sanderson, P.M., Jenkins, S.A., Leane, T.A., 2009. Interruptions and blood transfusion checks: lessons from the simulated operating room. Anesth. Analg. 108 (1), 219–222.
Lopetegui, M.A., Bai, S., Yen, P.-Y., Lai, A.M., Embi, P.J., Payne, P., 2013. Inter-Observer Reliability Assessments in Time Motion Studies: The Foundation for Meaningful Clinical Workflow Analysis. AMIA Annual Symposium, Washington, DC, USA.
Lopetegui, M., Yen, P.-Y., Lai, A., Jeffries, J., Embi, P., Payne, P., 2014. Time motion studies in healthcare: what are we talking about? J. Biomed. Inform. 49, 292–299.
Loukopoulos, L.D., Dismukes, R.K., Barshi, I., 2001. Cockpit interruptions and distractions: a line observation study. In: 11th International Symposium on Aviation Psychology. Columbus, Ohio.
Loukopoulos, L.D., Dismukes, R.K., Barshi, I., 2009. The Multitasking Myth: Handling Complexity in Real-World Operations. Ashgate Publishing Co., Surrey UK.
Mache, S., Vitzthum, K., Klapp, B.F., Groneberg, D.A., 2012. Doctors' working conditions in emergency care units in Germany: a real-time assessment. Emerg. Med. J. 29 (e1).
Mark, G., Gudith, D., Klocke, U., 2008. The cost of interrupted work: more speed and stress, Computer-Human Interactions. ACM, Florence, Italy, pp. 107–110.
Mark, G.J., Voida, S., Cardello, A.V., 2012. A pace not dictated by electrons: an empirical study of work without email. In: Konstan, J.A. (Ed.), CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Austin, Texas.
McFarlane, D.C., 1997. Interruption of People in Human-Computer Interaction: A General Unifying Definition of Human Interruption and Taxonomy. NRL-FR-5510-97-9870. Office of Naval Research, Arlington, VA, USA.
Mintzberg, H., 1970. Structured observation as a method to study managerial work. J. Manag. Stud. 7 (1), 87–104.
Nugus, P., Braithwaite, J., 2010. The dynamic interaction of quality and efficiency in the emergency department: squaring the circle? Soc. Sci. Med. 70 (4), 511–517.
Pollitt, L.C., Reece, S.E., Mideo, N., Nussey, D.H., Colegrave, N., 2012. The problem of auto-correlation in parasitology. PLoS Pathog. 8 (4), e1002590.
Rothman, K.J., Greenland, S., Poole, C., Lash, T.L., 2008. Causation and causal inference. In: Rothman, K.J., Greenland, S., Lash, T.L. (Eds.), Modern Epidemiology. Lippincott Williams & Wilkins, Philadelphia, USA.
Salvucci, D.D., Taatgen, N.A., Borst, J.P., 2009. Toward a unified theory of the multitasking continuum: from concurrent performance to task switching, interruption, and resumption. In: Proceedings of ACM CHI 2009 Conference on Human Factors in Computing Systems, ACM. Boston, MA, USA, pp. 1819–1828.
Sasangohar, F., Donmez, B., Trbovich, P., Easty, A.C., 2012. Not all interruptions are created equal: positive interruptions in healthcare. In: Human Factors and Ergonomics Society Annual Meeting, pp. 824–828.
Schildcrout, J.S., Heagerty, P.J., 2005. Regression analysis of longitudinal binary data with time-dependent environmental covariates: bias and efficiency. Biostatistics 6 (4), 633–652.
Schisterman, E.F., Cole, S.R., Platt, R.W., 2009. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology 20 (4), 488–495.
Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin Company, Boston, USA.
Sierhuis, M., Clancey, W.J., Van Hoof, R.J.J., 2007. Brahms: a multiagent modelling environment for simulating work processes and practices. Int. J. Simul. Process. Model. 3 (3), 134–152.
Strayer, D.L., Drews, F.A., 2006. Multi-tasking in the automobile. In: Kramer, A.F., Weigmann, D.A., Kirlik, A. (Eds.), Attention: From Theory to Practice. Oxford University Press, Oxford, UK.
Su, N.M., Mark, G.J., 2008. Communication Chains and Multitasking, CHI2008. ACM, Florence, Italy, pp. 83–92.
TimeCaT 3.9, 2013. TimeCaT Time Capture Tool ⟨www.timecat.org⟩. The Ohio State University.
Trafton, J.G., Altmann, E.M., Brock, D.P., Mintz, F.E., 2003. Preparing to resume an interrupted task: effects of prospective goal encoding and retrospective rehearsal. Int. J. Hum. Comput. Stud. 58 (5), 583–603.
Walter, S.R., Li, L., Dunsmuir, W.T.M., Westbrook, J.I., 2014. Managing competing demands through task-switching and multitasking: a multi-setting observational study of 200 clinicians over 1000 h. BMJ Qual. Saf. 23 (3), 231–241.
Watson, J.M., Strayer, D.L., 2010. Supertaskers: profiles in extraordinary multitasking ability. Psychon. Bull. Rev. 17 (4), 479–485.
Weigl, M., Hoffmann, F., Müller, A., Barth, N., Angerer, P., 2014. Hospital paediatricians' workflow interruptions, performance, and care quality: a unit-based controlled intervention. Eur. J. Pediatr. 173 (5), 637–645.
Weigl, M., Müller, A., Vincent, C., Angerer, P., Sevdalis, N., 2012. The association of workflow interruptions and hospital doctors' workload: a prospective observational study. BMJ Qual. Saf. 21 (5), 399–407.
Weigl, M., Müller, A., Zupanc, A., Glaser, J., Angerer, P., 2011. Hospital doctors' workflow interruptions and activities: an observation study. BMJ Qual. Saf. 20 (6), 491.
Westbrook, J.I., 2015. Interruptions and multi-tasking: moving the research agenda in new directions. BMJ Qual. Saf. http://dx.doi.org/10.1136/bmjqs-2014-003372, in press.
Westbrook, J.I., Ampt, A., 2009. Design, application and testing of the work observation method by activity timing (WOMBAT) to measure clinicians' patterns of work and communication. Int J Med Inf. 78, S25–S33.
Westbrook, J.I., Coiera, E., Dunsmuir, W.T.M., Brown, B.M., Kelk, N., Paoloni, R., Tran, C., 2010a. The impact of interruptions on clinical task completion. Qual. Saf. Health Care 19 (4), 284–289.
Westbrook, J.I., Woods, A., Rob, M.I., Dunsmuir, W., Day, R.O., 2010b. Association of interruptions with an increased risk and severity of medication administration errors. Arch. Intern. Med. 170 (8), 683–690.
Wickens, C.D., 2002. Multiple resources and performance prediction. Theor. Issues Ergon. Sci. 3 (2), 159–177.