Pre-print of: Lindley, S.E. and Monk, A.F. (2013). Measuring social behaviour as an indicator of experience. *Behaviour and Information Technology 32*, 10, 968-985. doi:10.1080/0144929X.2011.582148.

Measuring social behaviour as an indicator of experience

Siân E. Lindley and Andrew F. Monk

Department of Psychology
University of York
York, YO10 5DD, UK

Abstract

This paper explores and evaluates two techniques that measure aspects of social behaviour as an indicator of experience. The rationale driving the work is the idea that experience is entwined with social interaction and so, while experience itself is difficult to quantify, we might tap into it by measuring aspects of conversation that are related to it. Two techniques are considered as possible ways of doing this: (i) process measures of social behaviour derived from video analysis, and (ii) thin-slice ratings ascribed by naïve judges. Regarding (i), process measures of conversational equality, freedom and number of turns are shown to be reliable, sensitive, and linked to unfolding experience. Regarding (ii), a Thin Slice Enjoyment Scale is developed and shown to be a reliable and less time-consuming, but also less sensitive, alternative to the process measures. Both methods are of interest to researchers and practitioners who would wish to assess user experience in a group context. Additionally, analysis of the process measures is of broader relevance to researchers who conduct quantitative analyses of talk.

Keywords: Unfolding experience, empathic experience, process measure, quantitative, video analysis, thin slice.

1 Introduction

The field of Human-Computer Interaction (HCI) has undertaken a number of shifts in its recent history, two of the most prominent being the 'turn to the social' and the change of focus from usability to user experience, otherwise known as the second and third waves of HCI (Bødker, 2006). The associated evolution of methodologies, ideologies and analytical techniques is reflected in the current diversity of efforts to conceptualise, study and measure experience, with some researchers arguing that the very notion of trying to measure experience is flawed, while others propose that they have devised objective and quantifiable ways of doing just that.

In this paper we highlight and analyse two techniques that might be understood as offering quantifiable ways of tapping into experience. We argue that, in some situations at least, experience is interlinked with social interaction. Accordingly, we explore the question of whether certain quantifiable aspects of social interaction might serve as indicators of unfolding experience. The first technique involves taking various process measures of conversation through a video analysis of groups of

friends who are talking about their photos. Photograph sharing is undertaken in two experimental conditions, which offer different affordances for social interaction (cf. Gaver, 1996) and that were hypothesised (and have been shown to) present different possibilities for experience. The second technique draws on third-party, empathised, ratings. This approach to the rating of 'thin slices' of behaviour is well-established in psychological research, but does not feature in the field of HCI. In this method, indicators of experience are not defined by the researchers, but are instead intrinsic to scores ascribed by naïve judges.

The principal aim of the paper is to make a methodological contribution. We consider the reliability and sensitivity of the various measures used, and introduce a rating scale that may be used as a low-effort but nevertheless reliable way of gathering quantitative data from video. Further to this, we hope to offer a more considered exploration of the measures we describe, with the aim of examining how meaningfully they can be used as indicators of experience and social interaction. Therefore, a further aim of the paper is to provide a resource for researchers and practitioners who are interested in using video analysis as a way of measuring specific aspects of social interaction, or are interested in their potential to serve as indicators of user experience.

The outline of this paper is as follows. We will describe why these sets of measures were chosen by drawing on a review of the related literature, and specify how they were derived. Having detailed the methods, we will explore the sensitivity of the various measures to experimental manipulations and examine their inter-relationships. We will finish with a discussion of what the different measures are indicative of, both in terms of social interaction and experience.

1.1 Experience in HCI

We have already alluded to the fact that although a need to consider experience is widely accepted in HCI, attempts to measure it as a quantifiable variable are not without criticism. We will therefore begin this literature review by acknowledging the debate surrounding these issues, starting by highlighting different approaches to conceptualising experience and emotion[1].

Particularly noteworthy here is work by Boehner and colleagues (2007) who suggest that, just as cognition has been reconceptualised in the field of HCI from an internal phenomenon to one that is distributed, so should emotion be placed in a broader context. Far from being seen as a delineated and objective biological actuality, here

[1] Following Scherer (2005), we define emotion as entwining subjective feeling with action, expression and physical sensations, as being anchored in events and therefore subject to rapid change, and as being distinct from more enduring moods, traits, preferences or attitudes. This includes states such as interest, excitement, boredom, and impatience, which overlap with the experiences in which we are interested. While we use the term 'experience' to be consistent with our previous work, we also address the literature on 'emotion', and especially that relating to its conceptualisation and measurement.

emotion is positioned as a social and cultural product, experienced through encounters with others. According to this view, interactive systems should be designed to help people better understand and experience their emotions, and in evaluating them we should steer clear of attempts at direct measurement. This perspective has been extended by Höök et al. (2008), who emphasise the importance of embodied experience, and propose that the way in which we make sense of emotions is a combination of experiential processes in our bodies and the ways in which they are expressed in context and through interactions with others. They note that in interaction design, the meaning of a system is completed by the person experiencing it, thus individual experience cannot be separated from an overall experience that arises through a dialogue between that individual and other people or technologies. However, Höök et al. do suggest that by closing off experience as 'ineffable', Boehner et al. are at risk of mystifying it. In contrast, they argue that it is possible to speak about the qualities of experiences without reducing them to something less than the original. However, both would probably view the approach that we will outline in this paper as reductionist. In the next section, we describe our rationale for exploring such an approach.

1.2 Rationale of Assessing Experience as it Unfolds

In our own previous work (Lindley & Monk, 2008) we have attempted to use the social behaviour of collocated groups as a way of tapping into experience. The inspiration for this was drawn from a distinction made by McCarthy and Wright (2004) between experience as it unfolds and that which is recounted. Here, recounting is characterised as the telling of an experience either to oneself or to others after it has occurred, in contrast to various other processes that allow people to make sense of their experiences as they are happening (these include anticipating, connecting, interpreting, reflecting, and appropriating). McCarthy and Wright suggest that recounting is part of, and indeed alters, experience; thus experiences are remade each time they are revisited, and events, at least as experienced initially, may differ from retrospective views of those same events. In this holistic account, people are understood to continuously make sense of their experiences in ways that are uniquely personal to them.

While McCarthy and Wright's (2004) sense-making processes are intertwined, we suggest that if experience is understood as developing and altering over time, there is a need within HCI to tap into it as it unfolds. Furthermore, we suggest that this is especially the case when designing technology for recreation or leisure (for example, in game-play), in which the experience of actually interacting with a system is key. This requires a rather different approach to that often adopted within HCI, where rating scales are used as a way of evaluating an experience after it has occurred (see e.g., Tullis & Albert, 2008; Rodden et al., 2010). We argue that rating scales are a form of recounting, and that it is therefore important to complement such methods with ways of tapping into experiences as they happen. This is not to suggest that later accounts of the same experience are not relevant. Rather, we are proposing that, because experience as it happens can be somewhat different to that which is integrated into one's personal history, there are occasions where it is important to access that experience in the moment. However, the process of doing so raises a

number of difficulties; most notably, it highlights the question of how we can tap into unfolding experience without altering it. For instance, experience sampling (Csikszentmihalyi & Rathunde, 1993; see also Schleicher & Tröster, 2009, for a PC-based implementation to quantify 'joy-of-use'), where informants are interrupted briefly at arbitrarily defined moments to give ratings, may work in some circumstances but has the potential to disrupt a social experience. Similarly, it is difficult to believe that one's memory of an experience is the same at the start as it is at the end of rating the 36 items in Jackson and Marsh's (1996) Flow State Scale.

We suggest that understanding experience as it is happening might be achieved by identifying process measures of social behaviour, which are closely entwined with experience. Thus, while our approach is very different to that taken by Boehner et al. (2007) and Höök et al. (2008), we are in agreement that experience is bound up with social interaction, and we propose further that this relationship could be a means of understanding experience as it unfolds. Our previous paper (Lindley & Monk, 2008) was an exploration of how various affordances for social interaction might influence experience during photograph sharing. Following Gaver's (1996) proposal that the design of technology influences the nature of social interactions unfolding around it, we predicted that social behaviour occurring around particular configurations of technology would be influenced by the way in which that technology was set up. Furthermore, we suggested that this influence on social interaction would have a resultant effect on experience, and that this changing experience would in turn influence social behaviour. Consequently, measures of social behaviour may in part reflect unfolding experience. In the current paper we explore and evaluate this proposal for a variety of measures of group process, some not included in our previous paper (Lindley & Monk, 2008). Before we present further details of this approach, we give an overview of other approaches to the continuous measurement of experience in HCI.

1.3 Continuous Measurement of Experience in HCI

Of course, we are not the first to attempt to tap into experience as it unfolds, and other measures have been proposed that are not linked to social behaviour. Most notably, Mandryk et al. (2005) have sought to address the lack of a quantitative and objective means for assessing technologies designed to support game-based, playful experiences through the modelling of emotion using physiological measures, such as the galvanic skin response. Shastri et al. (2010) have also used physiological measures, combined with a questionnaire, to gauge mental engagement and enjoyment during a monitoring task, and researchers in the field of affective computing have utilised physiological data so as to design computers that can interpret the experience of the user and respond accordingly (e.g. Picard, 1997).

A rather different approach is taken by Isbister et al. (2006), who developed the Sensual Evaluation Instrument (SEI) as a way for people to express their emotions by selecting from and handling a variety of sculpted objects during game play. The SEI was developed in line with arguments later expressed by Höök et al. (2008), in that the findings are used as a cue for further discussion and reflection rather than serving as a form of output in themselves. Wright and McCarthy (2008) take a similar

view, arguing for techniques such as cultural probes (Gaver & Dunne, 1999), which open up a dialogue with the user and encourage researchers to use empathy in reaching an understanding of their experience. An approach that combines these two methods is presented by Burmester et al. (2010), who describe a valence method that can be used during formative evaluations of interactive products. Users press one of two buttons to show when they are feeling positive or negative whilst exploring an interface, with the valence markers then being revisited in a retrospective interview. A final device worth mentioning, which also resonates with work on embodied emotion, is the emotion slider, developed by Laurans et al. (2009). Here it was shown that participants reacted more quickly when pushing to indicate positive valence and pulling to indicate negative valence than vice versa.

In addition to approaches such as these, researchers use video analysis as a way of studying interactions as they unfold (e.g. Scott et al., 2003). Because of our focus on social interaction as an indicator of experience, this is the approach that we take here. In the next section, we outline the reasons behind our own selection of dependent variables.

1.4 Process Measures from Conversation Analysis

The choice of process measures presented here was primarily influenced by Edelsky's (1981) work, which used conversation analysis to examine social interaction during a series of meetings. She observed two contrasting styles of conversation, which have since been termed *cooperative floors* and *exclusive floors* (Morgenthaler, 1990). The cooperative floor is typified by a feeling of participants being "on the same wavelength" in a conversation that is a "free-for-all" (Edelsky, 1981, p. 384); here there is a sense that no one owns the floor, it being perfectly acceptable for everyone to talk at once. In contrast, the exclusive floor is characterised by a sense of orderliness, with only one person owning the floor at a time and turns rarely overlapping. The cooperative floor seems to capture the sense of cohesiveness and engagement that is associated with positive experiences in recreational situations. Furthermore, these characteristics of talk have since been linked with informal social interaction by many researchers (Coates, 1988; Dunne & Ng, 1994; Tannen, 1984) and with enjoyment by others (Monk & Reed, 2007).

On the basis of this previous work, we identified a number of quantitative measures that might draw on these aspects of conversation. In particular, measures of conversational equality and freedom (or interactivity, following Carletta, Garrod & Fraser-Krauss, 1998) seem to resonate particularly strongly with Edelsky's (1981) description, as do measures of conversational fluency through the occurrence of frequent turns (cf. Daly-Jones et al., 1998) and turn overlap. We have previously used these measures as dependent variables (Lindley & Monk, 2008), and in this paper we examine in more depth their sensitivity and validity, along with other measures of mean turn duration (also used by Daly-Jones et al.) and turn synchronisation (following Watts, Monk, & Daly-Jones, 1996).

It is worth noting that these process measures are also useful as a way of investigating social behaviour independent of experience, and that this is quite

common in HCI. Indeed, process measures often make a good complement, or alternative, to other, more typical usability measures, such as those relating to performance. This is particularly the case when designing for recreational situations, where traditional performance measures (such as completion time or number of errors) become less relevant, and the process of *doing* an activity is as important as the end result. Limitations of this approach include its time-consuming nature and the need to specify certain behaviours to focus on. Thus, in addition to the video analysis used in our own work, we also explored an alternative way of gathering data from video, through ratings of 'thin slices' of behaviour.

1.5 Rating Thin Slices of Behaviour

Rating thin slices of video data is fairly standard in the psychology literature, and tends to involve asking judges to watch short excerpts of social behaviour and to draw inferences from these. The procedure has been used to explore stable variables such as personality traits, intelligence and sexual preference, along with changing states such as the presence or absence of deception (see Ambady, Bernieri, & Richeson, 2000, for an overview), and has been shown to be predictive of factors such as teacher competence (Ambady & Rosenthal, 1993). Research has demonstrated that people are surprisingly good at making these types of judgement, even when basing them on slices of behaviour that are as 'thin' as 10 seconds in length. Ambady and Rosenthal (1992) suggest that this is because these inferences are made through the subconscious decoding of expressive behaviour; a "gestalt, molar impression" (p. 439; Ambady & Rosenthal, 1993) is quickly derived, without any specific aspect of behaviour being particularly predictive. Researchers often aim to encourage the formation of this gestalt impression by reducing the amount of information available to raters, for example by removing speech content while retaining tone of voice, or extinguishing facial expressions while preserving gross movements (Bernieri, Davis, Rosenthal, & Knee, 1994). In their meta-analysis of such studies, Ambady and Rosenthal (1992) found that judgement accuracy does not notably increase with the inclusion of information about speech content or with exposure to longer slices (although see Carney, Colvin, & Hall (2007) for evidence that 60 seconds is optimal when measuring positive affect, extraversion and agreeableness).

Ambady and Rosenthal (1992) propose that thin slice judgements can be usefully made so long as the variables in question are observable and there is an affective or interpersonal component. Furthermore, although the vast majority of research in this area focuses on person-specific constructs, there are examples of thin slice research that explore aspects of social interaction. For example, Grahe and Berniere (1999) have used ratings of thin slices of video content to explore rapport, a construct that is particularly relevant to the present paper and that resonates with the notion of the cooperative floor (Edelsky, 1981) described above. Grahe and Berniere suggest that when judging rapport, nonverbal behaviour (note that this includes tone of voice) is essential, and claim that their most accurate judges relied on interactional synchrony (or coordination of behaviours), proximity of participants, and expressivity (or animation of behaviour) in making their ratings. In the current paper, we will use ratings of thin slices to explore both experience and social

behaviour. As the process measures derived from our video analysis pertain almost exclusively to the characteristics of conversation, and neglect the types of behaviour highlighted by Grahe and Berniere, it will be interesting to see how closely the two sets of data are aligned, and in particular, to explore whether the process measures suggested in our previous paper correspond with raters' empathic judgements. In the next section, we will describe in detail the methods for deriving these two sets of data.

2 Methods

2.1 The Video Data

The video data used in the analyses presented here were part of a corpus collected during the first author's PhD research into affordances for social interaction (Lindley, 2006). The experiment in question and its results are not the focus of this paper and have been published elsewhere (Lindley & Monk, 2008), but we will now briefly describe the design for purposes of clarity. The experiment was a within-groups manipulation of interpersonal awareness during photograph sharing: eight groups of three friends talked about their photos in two seating configurations, which were designed to mimic the experience of 'hovering' behind someone using a PC, versus 'huddling' around a set of photos. To be specific, the independent variable was interpersonal awareness, with two levels, high awareness and low awareness. This manipulation was expected to offer different experiences to the group members across the two conditions, and indeed, the dependent variables relating to aspects of conversation and subjective ratings of experience suggested that this was the case. Specifically, in the high interpersonal awareness condition, the conversation was found to be more equal and higher in freedom[2] as well as being rated as more enjoyable.[3]

Additionally, because photo-talk has been shown to differ strongly according to the type of photos that are being discussed (Frohlich, Kuchinsky, Pering, Don, & Ariss, 2002), participants were asked to share equal numbers of photos supporting reminiscing photo-talk (occurring when photos depict events that everyone present attended) and storytelling photo-talk (occurring when photos depict events at which only the photographer was present). This allowed for analysis of the effects of a second independent variable, that of photograph content. Process measures of

[2] The process for deriving group scores of conversational equality and freedom will be described in more detail in Section 2.2. Analysis of these scores using Wilcoxon matched-pairs signed-ranks tests showed differences to be statistically significant ($p < .05$ in both cases), as reported by Lindley and Monk (2008).

[3] Participants rated their agreement with photo sharing as 'enjoyable' and as 'unfavourable' on a 5-point scale. The latter was subtracted from the former and these composite scores were then averaged across participants in each group. A two-tailed within-subjects t-test showed differences to reach statistical significance ($p < .01$), as reported by Lindley and Monk (2008).

conversation were found to differ in accordance with Frohlich et al.'s prior conversation analysis of photo-talk: conversational equality, freedom and fluency were all found to be higher during reminiscing than storytelling[4]. Furthermore, the effect of the experimental manipulation was found to be much stronger during storytelling than when the groups were reminiscing (Lindley & Monk, 2008).

The video data that we will analyse in the current paper are of eight groups in each of four conditions: high awareness during reminiscing, high awareness during storytelling, low awareness during reminiscing and low awareness during storytelling. This results in 32 segments of video. The last four minutes of each video segment were coded by the first author and rated by naïve judges for the thin slice study. We focused on the last four minutes of the conversation to reduce effects of the novelty of the situation on the behaviour sampled. The two categories of dependent variable are detailed further in the next two sections.

2.2 Measures of Group Process

The coding scheme that underpins the data analysis presented here was conducted using The Observer© and focused on conversational turns and smiling. Smiles were coded as 'events' (according to the terminology used The Observer©) and were recorded through a single timestamp. Coding the conversation was more complicated, with turns, other utterances, and periods of silence all being recorded as a continuous record of behaviour for each participant (See Figure 1 for an example of a time-event plot of the coding scheme). Turns were defined as a period of meaningful talk that contributed to the group's conversation, while other utterances included backchannels (Yngve, 1970), defined as verbalisations that signal continuing attention (e.g., saying "mmm" or "yeah"), and laughter. Only turns are used in the data presented here. They were coded as 'states', with a first timestamp for their onset and a second for their offset, allowing durations to be analysed. The offset of a turn was defined as the point at which the speaker stops talking in such a way as to allow somebody else to take the floor, even if nobody else does.

[Figure 1 here]

These records were analysed to produce the summary statistics alluded to in the literature review. As already mentioned, measures of conversational equality, freedom, number of turns, and turn overlap have been reported in our previous paper (Lindley & Monk, 2008). We will summarise these here without repeating the full details of how they are calculated. The details for calculating mean turn duration and turn synchronisation, which were not included in our previous publication, are presented in full.

---

[4] Scores of conversational equality and freedom were analysed using Wilcoxon matched-pairs signed-ranks tests ($p < .05$ in both cases). Indicators of conversational fluency were analysed using two-way within-subjects ANOVAs, which showed significant effects for the number of turns ($p < .01$) and mean turn overlap ($p < .01$), as reported by Lindley and Monk (2008).

Conversational Equality and Freedom

Conversational equality is a measure based on the amount of time that individuals within a group talk for, and how evenly this is distributed within the group. In the data analyses presented here, equality is based on conversational turns (to the exclusion of other utterances, such as backchannels and laughter). Our method for calculating equality is based on that proposed by Carletta et al. (1998) and produces a score between 0 and 1, where 1 represents a group in which all members talk for the same length of time in the four minutes of video that were analysed.

The measure of conversational freedom (also following Carletta et al., 1998) is derived from patterns of turn taking within a group, and indicates how frequently individuals within it take turns in the conversation immediately after specific other group members. For example, if person W only speaks after person V and never after person X, freedom will be low. Again, in our data, freedom is based on turns and not on other utterances, and produces a score between 0 and 1.

Number of Turns and Mean Turn Duration

In our previous paper, a measure of the number of turns in the conversation (summed over all group members during the four minutes of video that were analysed) was used as an indicator of conversational fluency. In this paper we also consider the related measure of the mean duration of these turns. Number of turns and mean turn duration are often used concurrently and are expected to be strongly inversely proportional for any one conversation; if there are many conversational turns during a four minute period, it follows that these turns will be short. However, the potential for periods of mutual silence and overlapping talk means that a correlation between the two scores is unlikely to be perfect and so, given the exploratory nature of this paper, it seems worth considering both. Conversations with a large number of short turns are taken to be more fluent than those featuring a lesser number of longer turns.

Turn Overlap and Turn Synchronisation

Turn overlap, measured as the amount of time that group members engage in overlapping turns, is often taken as an indicator of low conversational fluency, and indeed, groups that are unable to coordinate turn taking may experience overlapping speech as disruptive or frustrating. However, in Edelsky's (1981) description of the cooperative floor, it is emphasised that the floor belongs to no one, and that everyone may talk at once. Because of this uncertainty, we take the opportunity in this paper to explore the meaning of this measure both in terms of social interaction and experience. Furthermore, we consider it alongside an additional measure, that of turn synchronisation (Watts et al., 1996). This is used here because turn overlap has the potential to be confounded with the overall amount of conversation; when groups talk more there is more potential for this talk to occur simultaneously.

To avoid this potential pitfall, measures of turn synchronisation take into account the amount of time that two partners in a conversation might speak simultaneously by chance. The expected chance proportion of overlapping speech is calculated as a product of the proportion of time that each of two partners is speaking: if person V

speaks for .3 of the time and person W also speaks for .3 of the time, and both ignore what the other is doing, then turns can be expected to overlap .09 of the time. However, in normal conversation, V and W will follow social conventions pertaining to turn-taking rules, and thus the proportion of overlapping speech will be reduced. Turn synchronisation is calculated as the expected proportion of overlap minus the observed proportion[5], and so a higher score indicates a more coordinated conversation. Overlap and synchronisation thus reflect potentially different aspects of the conversation and synchronisation is to be preferred as a measure of coordination that is not confounded with the amount of time that group members spend talking.

In the video that was coded, instances in which all three group members spoke simultaneously were exceedingly rare. Consequently, for both turn overlap and turn synchronisation three pairwise scores are computed for each group. Pairs were identified according to two roles that are inherent in photograph sharing (cf. Frohlich et al., 2002): the photographer (the person who provides the photos) and the audience (the people who the photos are being shown to). For all experimental conditions, three pairwise scores were calculated: (i) between the photographer and the audience member who spoke most (P and A_{Max}), (ii) between the photographer and the audience member who spoke least (P and A_{Min}), and (iii) between the two audience members (A_{Max} and A_{Min}).

Number of Smiles

Finally, the number of smiles seen during the four minutes of conversation was counted for each condition to provide a measure of behaviour unrelated to verbalisations but potentially related to experience.

Reliability of the Process Measures

Before considering these measures further, it is necessary to demonstrate that they are reliable. In addition to the coding performed by the first author, an additional coder was trained to use The Observer© and to identify the different behaviours underlying the coding scheme. He went on to code 16 video clips for the purposes of calculating inter-rater reliability. The first author was not present when he did this to avoid influencing his analysis, although an instruction sheet was available for him to refer back to. Table 1 gives the inter-rater reliability coefficients obtained by correlating the measures extracted from the second coder's time-stamped record with those obtained from that generated by the first author. Each of these reliability coefficients are Pearson correlations computed from a particular score as derived from the coding of the first author and the additional coder respectively, i.e., all these scores are derived from the same two sets of time-stamped records. In this context, a measure of inter-rater reliability is preferred to a measure of inter-rater agreement that directly compares individual time-stamped records (e.g., Cohen's Kappa) as the reliability represents a correlation against which all subsequent

[5] Note that Watts and Monk (1996) actually proposed observed frequency minus the expected frequency as a measure of synchronisation for gaze, but the alternative given here makes more sense for describing the coordination of speech.

correlations with that variable may be judged. Note that all the correlations reported in this paper are Pearson correlations (r) where the sampling unit is video clip, or partial correlations controlling for experimental condition and photograph content (see Section 3.2 for an explanation of the latter).

[Table 1 here]

It is generally accepted that reliability coefficients of .80 and above indicate scores that are highly reliable (cf. Baesler & Burgoon, 1987), and the analysis here shows that all of the reliability coefficients are high except that for the number of smiles. This may in part be because smiling is easier to miss; smiles are not audible and can be fleeting. Additionally, it is sometimes hard to evaluate where one smile has ended and another has begun, making it difficult to decide whether to code a period of smiling as only one lasting smile or as a number of smiles. The moderate coefficient of .64 suggests that it is acceptable to use the number of smiles for the purpose of comparing means across experimental conditions, although it puts a severe limit on the maximum correlation that it might have with any other variable. To allow this measure to be considered with the others it will be included in the analysis of the results, but given less weight than statistics based on conversational turn-taking.

2.3 Ratings of Thin Slices by Naïve Judges

The four minute video segments used to derive the process measures were also used to elicit ratings from naïve judges. Formal recommendations for the sampling of thin slices are absent from the literature, but common strategies include: taking three samples, from the beginning, middle and end of a behavioural stream; choosing an arbitrary time-stamp at which to begin; and choosing a meaningful behavioural event, an approach that allows duration to vary (Ambady, Bernieri & Richeson, 2000). For the purposes of this paper we wished to support a direct comparison between thin slice ratings and the process measures, and so we used the same video clips to derive both sets of data. Eight judges participated in the rating experiment, comprising four males and four females, with a mean age of 19.83 years (standard deviation 2.36 years). All of the judges were first year undergraduate students from the University of York. They did not know any of the people who appeared in the videos that they rated and were paid £15 each for their contribution.

Each judge attended three one-hour sessions. During the first session they were asked to sign a consent and confidentiality form and then watch and rate eight video clips. The eight clips consisted of one clip of each group that participated in the experiment described above, with two clips representing each of the different combinations of high awareness/reminiscing, high awareness/storytelling, low awareness/reminiscing and low awareness/storytelling. The order of the clips was counterbalanced across the eight judges, and each of the 32 clips was viewed twice (by two separate judges) during the first session, so that the results for all clips would be equally affected by scores from judges who had not yet become accustomed to using the rating scale. In the following two sessions, the judges watched and rated the remaining 24 video clips (12 per session) in a pseudo-random order. No more than two clips from the same group were viewed by any judge in the

same session. Clips from the same group were not viewed without at least three other video clips appearing between them.

A rating scale with nine items was used for the judges to score the groups' behaviour and is given in Appendix A. The items were divided into two sections: (a) experience, consisting of ratings of: enjoyment, fun, formality, naturalness and absorbedness, and (b) conversation, consisting of ratings of the conversation as: equal, a free-for-all, flowing and fluent. Ratings were made on a 10 cm analogue scale.

The 'Thin Slice Enjoyment Scale'

At this point, it is worth reporting the internal consistency (or reliability) of the rating scale and performing an exploratory factor analysis of these results. These two analyses show that firstly, the eight judges responded to the video data in a very similar way, making it possible to average across their scores, and secondly, that only one factor underlies their responses, meaning that the nine items on the rating scale can be used to produce a single score. This makes all further analyses much simpler, and so the details of these two analyses will now be briefly presented. This section may be of particular interest to researchers who are considering using the rating scale in analyses of their own.

Reliability of the Rating Scale

For each of the nine rating scale items, Cronbach's α was calculated to assess the consistency of ratings across the eight judges (treated as item in the analysis) and 32 video clips (treated as sampling unit). Table 2 gives these reliability coefficients for each item. The values are uniformly high, making it acceptable to simply average across judges so as to provide a score for each video clip for all subsequent analyses.

[Table 2 here]

Factor Analysis

Following this check of the internal consistency of the data, a Principal Components Analysis was conducted. This indicates that there is only one factor underlying the judges' responses to the different scales: only one component has an eigenvalue that exceeds the value of one, and the slope of the Scree Plot does not change direction beyond the first component (cf. Kline, 1994). Using Principal Axis Factoring to derive the optimal factor solution gives the factor loadings shown in Table 3.

[Table 3 here]

The four strongest weighted items in this factor solution are Enjoyment, Conversation as fluent, Fun, and Conversation as flowing, suggesting that the judges view enjoyment and conversational fluency as being related; indeed, there is no apparent distinction between scores for experience-related items and scores for conversation-related items, despite the fact that these were presented in two separate sections on the rating form. Given the resulting single factor structure, it makes sense to simplify the data by using a composite score calculated by summing the mean ratings (across judges) for each scale item (with the score for formality being reverse-coded). This composite score will be referred to as *empathised enjoyment*. Analysis of the internal consistency of this scale shows that Cronbach's α

is .97 and cannot be improved by removing any of the scale items.

3 The Dependent Variables

Having outlined the process measures and described how the measure of empathised enjoyment was derived from thin slice ratings, we are in a position to explore the sensitivity of these various dependent variables to experimental manipulations, and their inter-relationships. As a way of orienting the reader as we do so, a summary of all of the measures discussed so far is given in Table 4.

[Table 4 here]

3.1 Sensitivity of the Dependent Variables

The sensitivity of the dependent variables to experimental manipulations can be explored by measuring their effect sizes along with the level of statistical significance associated with each difference, so as to gain an insight into how effective each one is likely to be at picking up differences in group behaviour. It is worth highlighting again that our participants did rate their experience as being significantly different across the manipulation of peripheral awareness (comparisons of ratings were not made for photograph content; see Lindley & Monk, 2008), thus if the variables are to prove useful then reasonable effect sizes would be required here. However, it should be noted that due to the small sample size, the data is intended to allow for an exploration of trends, as opposed to being the basis for firm conclusions. Trends in the data will be interpreted according to Cohen's (1988) recommendation that an effect size of .10 represents a small effect, a value of .30 represents a medium effect and a value of .50 represents a large effect. Effect sizes for all measures are presented in Table 5.

[Table 5 here]

As reported in our previous paper (Lindley & Monk, 2008), process measures of conversational equality and freedom show significant differences across the experimental manipulation of peripheral awareness, while measures of the number of turns and turn overlap do not. In addition to this, we can also see from Table 5 that none of the other process measures of mean turn duration, turn synchronisation and number of smiles show differences that reach significance. However, it is important to note that according to Cohen's (1998) criteria, most of the effect sizes in Table 5 are large. This suggests that the process measures are sensitive to differences in group behaviour, and raises the possibility that with a sample size larger than 8, some of these differences would reach significance.

Further evidence for the sensitivity of the process measures is seen when looking at comparisons across photograph content. A broader range of significant results emerges here, and the direction of those differences is that which would be expected following Frohlich et al.'s (2002) conversation analysis of excerpts of photo-talk. In this prior work, reminiscing was noted as being more equal and as consisting of more overlapping and shorter turns than storytelling. This pattern is replicated here in an outcome that both complements Frohlich et al.'s previous

analysis and lends some validity to the current process measures.

It is also apparent that some of the measures that should be tapping into similar types of behaviour show similarities in terms of the magnitude of the effect sizes. Number of turns and mean turn duration show a similar pattern of results, as do measures of turn overlap and turn synchronisation. Of the latter two, it seems that overlap is the more sensitive, although this may be because direct measures of overlap are confounded with other differences in behaviour, such as the amount of talk. Furthermore, for both overlap and synchronisation, it is clear that there is some interplay between the pair under analysis and the way that behaviour differs across conditions; when the pair includes the photographer, effect sizes are smaller and less likely to reach significance than when only the audience members are analysed. We will return to a discussion of overlap, synchronisation and their meaning for different pairs later, but for now we will simply highlight the possibility that these measures are problematic to interpret.

Number of smiles is the only process measure not to show a significant difference for either manipulation, although it does show a large effect size when comparing across photograph content. In contrast, the newly proposed measure of empathised enjoyment shows large effects for both comparisons, and a highly significant difference across photograph content. As this measure is shown here to be potentially useful, in the next section we will explore further how it relates to the process measures, as well as investigating the inter-correlations between the process measures themselves.

3.2 Convergent and Discriminant Validity of the Dependent Variables

The above analysis of effect sizes indicates that the dependent variables are sensitive to the kinds of experimental manipulation we are interested in, but does not speak to their meaning. The process measures were selected as indicators of characteristics of conversation that might be markers of experience and sociability, but we have also noted that third-party judgements based on thin slices of behaviour might draw on other aspects of behaviour, such as behavioural synchrony or eye contact. An exploration of the inter-correlations between these variables will therefore offer some insight into whether the process measures are actually linked to experience, as well as giving some insight into their effectiveness as measures of the characteristics of talk that we are interested in. It is worth noting that correlational analyses between thin slice ratings and process measures are often performed as a means of understanding which aspects of behaviour judges draw on when assigning their ratings. By studying these correlations, also known as cue dependencies (cf. Grahe & Berniere, 1999), we can assess how relevant the aspects of conversation that we have measured are when making judgements of enjoyment. Convergent validity, or patterns of high correlation, indicate related aspects of behaviour, while discriminant validity, or low correlations, indicate differing characteristics.

Correlations were computed between all the measures used in the analysis in the previous section. Two sets of correlations are presented in Table 6: those below the

diagonal are simple correlations (r) whereas those above the diagonal are partial correlations. The simple correlations include covariance due to the independent variables (low vs. high awareness and reminiscing vs. storytelling photograph content). Partial correlations were computed using two dummy variables to code for these manipulations and therefore eliminate these sources of covariance. The patterns of correlations and partial correlations are expected to be similar if there is covariance across naturally occurring variance in the groups. Where a simple correlation is present and the partial correlation absent, it would appear that the covariance observed is due to the coordinated effect of the experimental manipulations.

[Table 6 here]

Correlations with Empathised Enjoyment

Table 6 shows that the strongest correlations between the measure of empathised enjoyment and the process measures are for scores of conversational equality and freedom, suggesting that the judges relied more heavily on these aspects of conversation than others that we measured when assigning their ratings. Moderate partial correlations with number of turns as well as measures of turn overlap and turn synchronisation (when including the photographer in the pairwise analysis) suggest that these factors were also attended to, but to a lesser extent. It is worth highlighting also that the size of these correlations compares very favourably with the thin slice literature; for example, the overall effect size derived from Ambady and Rosenthal's (1992) meta-analysis was .39, with an estimated range of .34 to .48.

Surprisingly low partial correlations are seen for mean turn duration, turn synchronisation across the audience members, and the number of smiles. We will consider why this might be the case in the discussion, and in the meantime simply highlight that these process measures seem to be weaker indicators of experience than others. Before moving on to this discussion, we will explore the inter-relationships between the process measures themselves, in an effort to understand which aspects of conversation they are actually tapping into. We would expect, for example, all indicators of conversational fluency to show convergent validity, but we would not necessarily expect a relationship between conversational equality and freedom.

Correlations between Conversational Equality and Freedom

One of the most striking findings in Table 6 then, is the close relationship between measures of equality and freedom. In theory, these two variables are completely independent; for example, the speaker sequence {V, W, X, V, W, X} has the potential for equality to equal 1, if all participants speak for equal durations, but for freedom to equal 0, as W always speaks after V, X after W and V after X. To help understand why high correlations have in fact emerged, Figure 2 presents interaction diagrams of two conversations undertaken in different experimental conditions, but by the same group. The conversation depicted on the left has an equality score of .95 and a freedom score of .89, while that on the right has an equality score of .75 and a freedom score of .64.

[Figure 2 here]

If we look at the diagrams more closely, we can see that in both conversations one pair within the group dominates to a degree. However, this is much more exaggerated in the diagram on the right, in which X is almost completely sidelined. The consequence of X's seeming inability to find a way into the conversation is that both equality and freedom are reduced, giving some insight into why the two measures might be so tightly coupled in practice. This coupling may also reflect, to a certain extent, the nature of photo-talk. As demonstrated by Frohlich et al. (2002), photo-talk is associated with strong group roles, such as those of photographer and audience member. In many cases, and especially during storytelling photo-talk, it will be the photographer who drives the conversation. While a lack of equality is an obvious consequence of this, it will also result in a lack of conversational freedom: when the audience members do speak, their turns are more likely to follow those of the photographer. This highlights the importance of understanding the activity under study when using these process measures, and suggests that for photo-talk at least, measuring either conversational equality or freedom might be sufficient.

Correlations between Indicators of Conversational Fluency

Moving on to measures expected to reflect conversational fluency, some more predictable relationships emerge. As anticipated, number of turns and mean turn duration are inversely related (correlation -.77, partial correlation -.64) and both seem to be tapping into the same behaviour. Furthermore, number of turns correlates significantly with turn overlap, lending some initial support to the idea that overlapping turns might also be taken as an indicator of conversational fluency.

However, the relationship between overlap and number of turns is not repeated for synchronisation scores. Furthermore, and somewhat surprisingly, the three pairwise measures of turn overlap demonstrate discriminant validity with the three pairwise measures of synchronisation, i.e., despite the fact that synchronisation scores are derived from a measure of overlap, the two scores are measuring something different. Note also that turn synchronisation does not correlate with number of turns, demonstrating that turn overlap is linked to number of turns in a way that turn synchronisation is not. As noted earlier, turn overlap is likely to be confounded with the amount of talk (when people talk more their talk is more likely to overlap), and this raises the possibility that the correlation between number of turns and turn overlap may simply reflect this confound. Therefore, measures of turn overlap need to be interpreted carefully, and are not necessarily a good indicator of conversational fluency.

Having highlighted the limitations of measuring turn overlap, we can return to turn synchronisation, a variable that is carefully calculated to avoid being confounded with amount of talk. However, similar to the findings in Table 5, Table 6 indicates that interpreting the meaning of this measure is not without its problems. First of all, and unlike measures of overlap, the three pairwise synchronisation scores do not demonstrate convergent validity (i.e. they are not correlated with one another). This suggests that there is no sense that any one of the pairwise scores is indicative of coordination of conversation at the group level. Again, this pattern of results may be

related to the activity of photograph sharing. If the photographer is leading the conversation, then it is plausible that the two audience members will primarily try to coordinate their speech with the photographer, rather than with each other. Thus low synchronisation across the audience members will not necessarily be associated with low synchronisation across pairs that include the photographer.

Interestingly, it is the synchronisation scores that include the photographer that show convergent validity with process measures of equality and freedom, indicating that conversations that are more equal and free also show more coordination, at least with the photographer. When discussing the measures in section 2.2, we raised a question as to whether conversations with lots of overlap would be seen positively (in line with Edelsky's (1981) notion of the cooperative floor) or negatively. It seems that synchronised conversation is associated with some attributes of the cooperative floor, such as equality and freedom, but not with others, such as fluency.

Correlations with Number of Smiles

Other than synchronisation across the two audience members, number of smiles is the only other variable not to show any significant partial correlations with any of the other process measures. Obviously, this is the only process measure of behaviour other than talk, and it is also a variable with relatively lower reliability. However, its lack of correlation even with ratings of empathised enjoyment suggests that it needs to be interpreted carefully; these results show that smiling seems not to be a strong influence on perceived experience.

4 Discussion

This paper set out to explore various ways of measuring behaviour in a social context, which might serve as indicators of experience, and which also have relevance for studies of computer-mediated communication, computer-supported cooperative work and collaboration. Various process measures of group behaviour have been explored along with a new rating scale of empathised enjoyment. We will conclude by discussing in turn three elements of this work: (i) what the process measures tell us about different aspects of conversation, (ii) whether these can be interpreted as indicators of experience, and (iii) the potential utility of the newly proposed Thin Slice Enjoyment Scale.

4.1 Measures of Conversation

Beginning with the most straightforward of these issues, we have empirically examined a number of measures of group process which might be used in a range of contexts and all of which, except for the number of smiles, have high reliability. Whatever they are measuring, the variables are consistent and the scores obtained are more than just chance error. Furthermore, the variables are shown to be sensitive to an experimental manipulation that has elicited significantly different ratings by participants (see Lindley & Monk, 2008); all show at least one moderate effect size across the two experimental manipulations, although not all reveal significant differences. The analysis indicates that measures of equality, freedom, number of turns and mean turn duration are most sensitive to the experimental

manipulations. Equality and freedom show high convergent validity, indicating that they are tapping into similar aspects of conversation. Number of turns and mean turn duration are also strongly related, but these two measures are not highly correlated with equality and freedom. This suggests that indicators of conversational fluency such as number of turns and mean turn duration are not inextricably linked to conversational equality and freedom, and implies that researchers using video analysis to explore computer-mediated communication or computer-supported collaboration would be well advised to use at least one score from both sets of measures if they wish to get a balanced impression of the characteristics of a conversation. We also noted that the relationship between equality and freedom may reflect the nature of photo-talk, so depending on the group activity and research questions being addressed it may be useful to calculate both scores.

In contrast, some of the pairwise measures of turn overlap and synchronisation show smaller effect sizes for the manipulation of awareness, and fail to reach significance when comparing across photograph content. This point is worth highlighting, as the manipulation of photograph content in particular represents a rather radical manipulation of social context, which we would expect to have a strong effect on behaviour. Furthermore, the patterns of correlation for these variables underline the fact that turn overlap, but not synchronisation, is confounded with amount of talk. These findings suggest that measures of turn overlap in particular should be interpreted carefully, and might be best avoided as a dependent variable. Measures of synchronisation are potentially more useful, but share with overlap the complication of being pairwise as opposed to group scores. Nevertheless, there may be instances when researchers are specifically interested in the coordination of a conversation, in which case we would recommend that they measure turn synchronisation but use larger sample sizes than those analysed here.

Finally, the measure of smiling shows lower reliability and sensitivity than the other process measures, and also shows discriminant validity with them, i.e., it is measuring something different. The lower reliability of this measure may have had an impact on effect sizes and correlations, but it is also reasonable to suggest that measures of smiling are less sensitive to the experimental manipulations in question. The fact that smiling and measures of talk do not correlate would not normally be problematic; they are, after all, different aspects of social behaviour. However, this is an issue here if we are to claim that these variables are all indicators of positive experience. In the next section we consider which of the dependent variables that we have analysed might be used as indicators of unfolding experience, and which are of most use when simply treated as broader indicators of social interaction.

We conclude this section by recommending that researchers and practitioners who are interested in examining the characteristics of conversation, for example in the context of studying computer-mediated communication or collaboration, focus on measuring conversational equality, conversational freedom, and either number of turns or mean turn duration as an indicator of conversational fluency. Turn synchronisation may also be useful if understanding coordination is particularly pertinent, or if the focus is on pairs rather than groups.

4.2 Indicators of Unfolding Experience

At the outset of this paper, we highlighted Edelsky's (1981) concept of the cooperative floor as being the inspiration for our choice of process measures, and selected variables that might tap into equal, fluent and overlapping conversation. This line of reasoning receives some support from the finding that measures of conversational equality and freedom were drawn on most heavily by naïve judges when rating empathised enjoyment, while turn overlap is also an important cue. However, and somewhat surprisingly, increasing synchronisation (at least, when the photographer is involved) also correlates significantly with empathised enjoyment, i.e. coordinated conversations are rated as more enjoyable. This contrasts with the view of uncoordinated turn taking where anything goes as being indicative of positive unfolding experiences. One explanation for the fact that measures of both turn overlap and turn synchronisation correlate with empathised enjoyment can be found in the fact that overlap is confounded with amount of talk: it is certainly plausible that conversations featuring lots of talk (and consequently scoring high on overlap) would be rated positively by the naïve judges, while those lacking in coordination or showing breakdowns in turn taking (and consequently scoring low on synchronisation) would be rated negatively.

Notably, the naïve judges' ratings are not related to the amount of smiling, as measured through video analysis. There are a number of possible reasons for this beyond the lower reliability of this process measure: counting smiles neglects a great deal of information, including the apparent meaning of the smile, its genuineness and the context in which it occurs. When attending to experience, the judges may have implicitly responded to the extent of genuine or so-called *Duchenne* smiling (Ekman, Friesen, & O'Sullivan, 1988), or drawn on their knowledge of the wider social context. The fact that smiling can be used for many purposes, particularly in social situations, suggests that it is not a particularly good measure of unfolding experience. Indeed, it is possible that the occurrence of awkward smiling might explain the low but negative correlations in Table 6.

The above implies that when selecting indicators of unfolding experience, measures that identify conversations that are equal, free, coordinated and with a large number of turns are of most relevance. However, the lower sensitivity of synchronisation scores makes this measure more problematic.

4.3 Thin Slice Enjoyment Scale

A final outcome of this work is the suggestion that thin slice judgements of video extracts might be used as an alternative to the calculation of process measures. The labour involved in video analysis is considerable, and necessitates that the coder concentrate on specific aspects of group behaviour, at the expense of others. Using ratings of video clips, provided by naïve judges, is much less time-consuming. Furthermore, researchers could use shorter slices than those analysed here and fewer judges, as has been shown to be reliable and effective in the thin slice literature. The above discussion also cements how potentially difficult it is to make decisions regarding which aspects of behaviour to code and the dangers of not paying sufficient attention to the contexts in which they occur; many of the

variables that we measured are confounded with other aspects of behaviour or tap into experience in unexpected ways. In contrast, thin slice ratings are a natural way of integrating a wider range of visual and auditory qualities into the process of deriving dependent variables. While this line of research certainly does not fit with that advocated by Boehner, Höök and others, it does go some way to alleviating the problems they rightly note as being associated with attempts to objectify experience: in this case it is not necessary for the researcher to specify exactly which behavioural cues should be attended to.

This paper has presented a measure of empathised enjoyment that could be used as a dependent variable in its own right: the Thin Slice Enjoyment Scale is given in Appendix A. While further work is needed to cross-validate the scale using data independent of that used in its development, and to apply it in the analysis of contexts other than photo sharing, the version presented here has been shown to be reliable, reasonably sensitive, and to show convergent validity with other measures that, we argue, are indicative of positive unfolding experience. Furthermore, our analysis suggests other uses for thin slice ratings. In addition to providing scores of behaviour, these scores might also be used to indicate whether more in-depth video analysis is worth pursuing; if reasonable effect sizes are found when comparing thin slice ratings, larger effects might be expected through video analysis. Finally, thin slice ratings could inform researchers when choosing which behavioural cues to code: explorations of cue dependencies derived from an initial video analysis might be useful in indicating which aspects of behaviour (e.g. speech, gaze, body movement) to analyse in more depth.

4.4 Recounted Experience

Before concluding, we should also acknowledge some of the limitations of this work. In particular, readers may be wondering why no correlations have been presented using the ratings made by the participants in the experiment. As stated in the introduction, our argument for using social behaviour as an indicator of unfolding experience is motivated by the notion that this unfolding experience can only be measured while it is happening. Subjective measures such as rating scales form part of the process of recounting (albeit, only to oneself or to the experimenter), thus the distinction between process and subjective measures reflects differences between unfolding and recounted experience. Nevertheless, one would expect recounted experience to be influenced by the initial experience.

Because of this potential relationship, it would have been interesting to compare the measures of process and empathised enjoyment with measures of recounted experience. However, while the groups of friends who participated did complete a questionnaire after each experimental condition, separate questionnaires were not completed for the two different types of photo-talk. Therefore, while we were able to detect significant effects of the manipulation of social affordance represented by our experiment, the procedure was not designed to support a correlational analysis (in particular, *n* was not sufficiently large for conclusions to be drawn). We would suggest though that, as in our previous work, a combination of methods should be used when evaluating experience, so that both unfolding and recounted experience

can be explored. Indeed, following the argument that experiences are remade and reinterpreted as they are recounted (McCarthy & Wright, 2004), we do not claim that our methods measure any 'one true experience'. Instead, we suggest that they are indicative of unfolding experiences as they are encountered. Furthermore, we argue that they should be supplemented with and complemented by other methods to explore recounted experience, where there is more scope for methods drawing on reflection and richer forms of empathy. In both cases, it is important for the methodological context of the techniques used to be understood; when considering recounted experience, context will inevitably impact the nature of what is recounted.

4.5 Other Indicators of Experience

As a final remark, it is important to also acknowledge that measures derived from social interaction may not always be the most appropriate indicator of unfolding experience. The data analysed here is from a study of the effects of affordances for social interaction within a particular context, namely photo sharing, where interaction with other participants is the major activity involved. Similar process measures have been used as dependent variables when examining how interface design can affect the mediation of conversation between remote partners (Daly-Jones, Monk, & Watts, 1998; Finn, Sellen, & Wilbur, 1997), collaboration around shared interfaces (e.g., Billinghurst, Belcher, Gupta, & Kiyokawa, 2003), and levels of engagement when playing collaborative computer games (Lindley, Le Couteur, & Bianchi-Berthouze, 2008). These may represent rather different contexts to photo sharing, as attention is directed as much to interaction with a device as it is to interaction with other people (multi-player gaming is a good example of this). It would be extremely interesting to compare the effects of different manipulations relevant to these contexts using our measures as well as other indicators of continuous experience, such as those noted in Section 1.3. This could be a thought-provoking piece of follow-up work, allowing for a deeper consideration of where measures of social behaviour should sit within the wider research on indicators of experience.

5 Conclusion

In this paper we have analysed a number of measures of group process and have introduced the Thin Slice Enjoyment Scale. These variables are shown to be reliable and sensitive to experimental manipulations and we have highlighted certain of them as being indicative of unfolding experience. This viewpoint finds support from the face validity of the measures taken, their sensitivity to manipulations that have been rated as offering different experiences, and the divergent and convergent validity that they demonstrate. In particular, we recommend measures of conversational equality, freedom and number of turns as being indicative of unfolding experience, and suggest the use of thin slice ratings as an easier, but less sensitive, alternative.

Acknowledgements
We would like to thank Barry Hannon for his efforts with the video analysis, Simon Fletcher for help in setting up the equipment for the study, and Rainer Banse and Gerry Altmann for insightful comments.

Appendix A: Rating Scale for Empathised Enjoyment

Please rate each group's behaviour according to how you perceive them to feel or behave *as a whole* while sharing photographs together. Try not to imagine how *you* would feel in their position, but focus on how *they seem to feel* based on their behaviour.

Draw a mark through each line to indicate how much you agree with the corresponding statement.

For the first set of questions, focus on the *overall experience* of the group.

	Strongly disagree	*Strongly agree*
The photo sharing in this clip seemed formal.	_____	
The group in this clip are enjoying themselves.	_____	
The photo sharing in this clip seemed natural.	_____	
The photo sharing in this clip seemed to absorb the	_____	
The group in this clip are having fun.	_____	

For the next set of questions, pay particular attention to the group's *conversation*.

	Strongly disagree	*Strongly agree*
The conversation involves the whole group equally.	_____	
The group's conversation is flowing.	_____	
The group's conversation is a free-for-all.	_____	
The group's conversation is fluent.	_____	

References

Ambady, N., Bernieri, F., & Richeson, J. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In M. Zanna (Ed.), *Advances in Experimental Social Psychology*. London: Academic Press.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*(2), 256-274.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64*(3), 431-441.

Baesler, E. J., & Burgoon, J. K. (1987). Measurement and reliability of nonverbal behavior. *Journal of Nonverbal Behavior, 11*(4), 205-233.

Bernieri, F., Davis, J., Rosenthal, R., & Knee, C. (1994). Interactional synchrony and rapport: Measuring synchrony in displays devoid of sound and facial affect. *Personality and Social Psychology, 20*(3), 303-311.

Billinghurst, M., Belcher, D., Gupta, A., & Kiyokawa, K. (2003). Communication behaviors in colocated collaborative AR interfaces. *International Journal of Human-Computer Interaction, 16*(3), 395-423.

Bødker, S. (2006). When second wave HCI meets third wave challenges. In A. Mørch, K. Morgan, T. Bratteteig, G. Ghosh & D. Svanaes (Eds.), *Proceedings of the 4th Nordic conference on Human-Computer Interaction,* (pp. 1-8). ACM Press: New York.

Boehner, K., DePaula, R., Dourish, P. & Sengers, P. (2007). How emotion is made and measured. *International Journal of Human-Computer Studies, 65*, 275-291.

Burmester, M., Mast, M., Jäger, K. & Homans, H. (2010). Valence method for formative evaluation of user experience. In O.W. Bertelsen et al. (Eds.), *Proceedings of the 8th ACM conference on Designing Interactive Systems,* (pp. 364-367). ACM Press: New York.

Carletta, J., Garrod, S., & Fraser-Krauss, J. (1998). Communications and placement of authority in workplace groups: The consequences for innovation. *Small Group Research, 29*(5), 531-559.

Carney, D., Colvin, C., & Hall, J. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality, 41*, 1054-1072.

Coates, J. (1988). Gossip revisited: Language in all-female groups. In J. Coates & D. Cameron (Eds.), *Women in their speech communities* (pp. 94-122). London: Longham.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd edition)*. New York: Academic Press.

Csikszentmihalyi, M., & Rathunde, K. (1993). The measurement of flow in everyday life: towards a theory of emergent motivation. *Nebraska Symposium on Motivation, 40*, 57-97.

Daly-Jones, O., Monk, A., & Watts, L. (1998). Some advantages of video conferencing over high-quality audio conferencing: Fluency and awareness of attentional focus. *International Journal of Human-Computer Studies, 49*, 21-58.

Dunne, M., & Ng, S. H. (1994). Simultaneous speech in small group conversation: All-together-now and one-at-a-time? *Journal of Language and Social Psychology, 13*(1), 45-71.

Edelsky, C. (1981). Who's got the floor? *Language in Society, 10*, 381-421.

Ekman, P., Friesen, W. V., & O'Sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology, 54*, 414-420.

Finn, K. E., Sellen, A. J., & Wilbur, S. B. (1997). *Video-mediated communication*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Frohlich, D. M., Kuchinsky, A., Pering, C., Don, A., & Ariss, S. (2002). Requirements for photoware. In E. F. Churchill & J. McCarthy (Eds.), *Proceedings of the 2002 ACM conference on Computer supported cooperative work* (pp. 166-175). New York: ACM Press.

Gaver, W. W. (1996). Affordances for interaction: The social is material for design. *Ecological Psychology, 8*(2), 111-459.

Gaver, W. & Dunne, A. (1999). Projected realities: Conceptual design for cultural effect. In M.G. Williams & M.W. Altom (Eds.), *Proceedings of the 1999 SIGCHI conference on Human factors in computing systems* (pp. 600-607). New York: ACM Press.

Grahe, J., & Bernieri, F. (1999). The importance of nonverbal cues in judging rapport. *Journal of Nonverbal Behavior, 23*(4), 253-269.

Höök, K., Ståhl, A., Sundström, P. & Laaksolahti, J. (2008). Interactional empowerment. In M. Czerwinski, A. Lund & D. Tan (Eds.), *Proceedings of the 2008 SIGCHI conference on Human factors in computing systems* (pp. 647-656). New York: ACM Press.

Isbister, K., Höök, K., Sharp, M. & Laaksolahti, J. (2006). The sensual evaluation instrument: Developing an affective evaluation tool. In R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries & G. Olson (Eds.), *Proceedings of the 2006 SIGCHI conference on Human factors in computing systems*, (pp. 1163-1172). New York: ACM Press.

Jackson, S.A., & Marsh, H.W. (1996). Development and validation of a scale to measure optimal experience: The flow state scale. *Journal of Sport and Exercise Psychology, 18*, 17-35.

Kline, P. (1994). *An easy guide to factor analysis*. London: Routledge.

Laurans, G., Desmet, P.M.A. & Hekkert, P. (2009). The emotion slider: A self-report device for the continuous measurement of emotion. In *Proceedings of the 3rd international conference on Affective Computing and Intelligent Interaction* (pp. 1-6).

Lindley, S.E. (2006). *The effect of the affordances offered by shared interfaces on the social behaviour of collocated groups*. Ph.D. Thesis, University of York.

Lindley, S. E., Le Couteur, J., & Bianchi-Berthouze, N. (2008). Stirring up experience through movement in game play: Effects on engagement and social behaviour. In M. Czerwinski, A. Lund & D. Tan (Eds.), *Proceedings of the 2008 SIGCHI conference on Human factors in computing systems* (pp. 511-4). New York: ACM Press.

Lindley, S. E., & Monk, A. F. (2008). Social enjoyment with electronic photo displays: Awareness and control. *International Journal of Human Computer Studies, 66*(8), 587-604.

Mandryk, R. L., Inkpen, K. M., & Calvert, T. W. (2005). Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour and Information Technology, 25*(2), 141-158.

McCarthy, J., & Wright, P. (2004). *Technology as experience*. Cambridge, MA: MIT Press.

Monk, A. F., & Reed, D. J. (2007). Telephone conferences for fun: experimentation in people's homes. In A. Venkatesh, T. Gonzalvez, A. Monk & B. Buckner (Eds.), *Home informatics and telematics: ICT for the next billion. Proceedings of HOIT 2007* (pp. 201-204). New York: Springer.

Morgenthaler, L. (1990). A study of group process: Who's got what floor? *Journal of Pragmatics, 14*, 537-557.

Picard, R. W. (1997). *Affective Computing*. Cambridge, MA: MIT Press.

Rodden, K., Hutchinson, H., & Fu, X. (2010). Measuring the user experience on a large scale: User-centered metrics for web applications. In E. Mynatt & D. Schoner (Eds.), *Proceedings of the 2010 SIGCHI conference on Human factors in computing systems* (pp. 2395-2398). New York: ACM Press.

Scherer, K.R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44*(4), 695-729.

Schleicher, R. & Tröster, S. (2009). The 'Joy-Of-Use'-Button: Recording pleasant moments while using a PC. In T. Gross et al. (Eds.), *Proceedings of the 12th IFIP TC 13 international conference on Human-Computer Interaction, Part II* (pp. 630-633). Berlin: Springer-Verlag.

Scott, S.D., Mandryk, R.L., & Inkpen, K.M. (2003). Understanding children's collaborative interactions in shared environments. *Journal of Computer Assisted*

Learning, 19, 220-228.

Shastri, D., Fujiki, Y., Buffington, R., Tsimayrtzis, P., & Pavlidis, I. (2010). O job can you return my mojo? Improving human engagement and enjoyment in routine activities. In E. Mynatt & D. Schoner (Eds.), *Proceedings of the 2010 SIGCHI conference on Human factors in computing systems* (pp. 2491-2498). New York: ACM Press.

Tannen, D. (1984). *Conversational style: Analyzing talk among friends*. New Jersey: Ablex Publishing Corporation.

Tullis, T., & Albert, W. (2008). *Measuring the user experience: Collecting, analyzing and presenting usability metrics.* Morgan Kaufman.

Watts, L., Monk, A., & Daly-Jones, O. (1996). Inter-personal awareness and synchronization: Assessing the value of communication technologies. *International Journal of Human-Computer Studies, 44*, 849-873.

Wright, P. & McCarthy, J. (2008). Empathy and experience in HCI. In M. Czerwinski, A. Lund & D. Tan (Eds.), *Proceedings of the 2008 SIGCHI conference on Human factors in computing systems* (pp. 637-646). New York: ACM Press.

Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society.

Table 1. Inter-rater reliability correlations between process measures (N = 16 video clips).

	Inter-rater reliability (Pearson's *r*)
Equality of turns	.99
Freedom of turns	.92
Number of turns	.87
Mean turn duration	.91
Turn overlap for P and A_{Max}	.95
Turn overlap for P and A_{Min}	.95
Turn overlap for A_{Max} and A_{Min}	.95
Turn synchronisation for P and A_{Max}	.95
Turn synchronisation for P and A_{Min}	.95
Turn synchronisation for A_{Max} and A_{Min}	.87
Number of smiles	.64

Table 2. Internal consistency for each item on the rating scale, calculated from the scores provided by the eight judges (N = 32 video clips).

	Reliability for each item (Cronbach's α)
Judged enjoyment	.84
Judged fun	.85
Judged formality	.86
Judged naturalness	.77
Judged absorbedness	.80
Judged conversation as equal	.85
Judged conversation as a free-for-all	.77
Judged conversation as flowing	.77
Judged conversation as fluent	.73

Table 3. The rating scale items and their loadings on a single factor (N = 32 video clips).

	Factor loadings
Judged enjoyment	.97
Judged conversation as fluent	.96
Judged fun	.96
Judged conversation as flowing	.95
Judged naturalness	.90
Judged formality (R)	.87
Judged conversation as a free-for-all	.86
Judged absorbedness	.84
Judged conversation as equal	.77

R indicates reverse-coded items

Table 4. Summary of all dependent variables, as defined in Sections 2.2 and 2.3.

Dependent variable	Description
Equality of turns	How equally talk is distributed within the group, scored from 0 (conversation is dominated by one person) to 1 (everyone talks for equal durations).
Freedom of turns	How interactive turn taking is, scored from 0 (turn taking is predictable; each group member always speaks after the same other group member) to 1 (everyone is equally likely to speak after everyone else).
Number of turns	The number of turns during the four minutes of video that were analysed, summed across all group members. A large number of turns suggests high conversational fluency.
Mean turn duration	The mean duration of all turns. A short mean duration suggests high conversational fluency, and this should be inversely proportional to the number of turns, i.e. fluent conversations comprise many, short turns.
Turn overlap	The amount of time that turns overlap within a pair during the four minutes of video that were analysed. High levels of overlap have been associated with low conversational fluency but also with informal social chat.
Turn synchronisation	A score relating to turn overlap that is not confounded with the amount of talk (pure measures of overlap may be higher as the amount of talk increases). High synchronisation indicates a more coordinated conversation.
Number of smiles	The number of smiles summed across all group members.
Empathised enjoyment	Composite score based on thin slice ratings ascribed by naïve judges. A high score reflects higher ratings in terms of enjoyment and fluency of conversation.

Table 5. Effect sizes (r) of the process measures and of empathised enjoyment. Significant differences from analyses of variance are indicated as *$p < .05$, **$p < .01$, while a subscript indicates that Wilcoxon matched-pairs signed-rank tests were used to compare means. In this case the variable that was held constant is noted as follows: $reminiscing_{rem}$, $storytelling_{sto}$, $high_{hi}$, low_{lo}. In all cases, within-group comparisons are made for samples of size N = 8.

	Awareness	Photograph content
Equality of turns	$.64_{rem}$ $.84_{sto}$*	$.89_{hi}$* $.89_{lo}$*
Freedom of turns	$.45_{rem}$ $.74_{sto}$*	$.84_{hi}$* $.89_{lo}$*
Number of turns	.53	.91**
Mean turn duration	.60	.88**
Turn overlap for P and A_{Max}	.00	.86**
Turn overlap for P and A_{Min}	.18	.54
Turn overlap for A_{Max} and A_{Min}	$.40_{rem}$ $.49_{sto}$	$.84_{hi}$* $.89_{lo}$*
Turn synchronisation for P and A_{Max}	.25	.62
Turn synchronisation for P and A_{Min}	.43	.55
Turn synchronisation for A_{Max} and A_{Min}	.49	.96**
Number of smiles	.17	.59
Empathised enjoyment	.64	.93**

Table 6. Correlations between process measures (below the diagonal, df = 30) and partial correlations between process measures (above the diagonal, df = 28), the latter controlling for experimental condition and photograph content, *p <.05, **p <.01.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. Equality of turns		.80**	.25	-.12	.29	.30	.28	.59**	.57**	.01	.05	.62**
2. Freedom of turns	.88**		.21	.14	.33	.44*	.36	.53**	.50**	.02	-.09	.67**
3. Number of turns	.59**	.50**		-.64**	.80**	.52**	.64	-.02	-.12	-.16	-.14	.42*
4. Mean turn duration	-.56**	-.31	-.77**		-.29	.17	-.02	.33	.36	.04	-.13	.04
5. Turn overlap for P and A_{Max}	.50**	.49**	.84**	-.49**		.60**	.62**	-.01	.14	-.14	-.30	.53**
6. Turn overlap for P and A_{Min}	.31	.42*	.53**	.01	.61**		.72**	.35	.18	-.17	-.31	.60**
7. Turn overlap for A_{Max} and A_{Min}	.60**	.60**	.76**	-.36*	.71**	.68**		.28	.31	-.26	-.28	.49**
8. Turn synchronisation for P and A_{Max}	.16	.28	-.14	.38*	-.14	.27	.11		-.33	.26	-.04	.40*
9. Turn synchronisation for P and A_{Min}	.58**	.58**	.11	.05	.28	.23	.44*	-.23		.02	.08	.47**
10. Turn synchronisation for A_{Max} and A_{Min}	.68**	.53**	.40*	-.48**	.31	.07	.35*	.32	-.27		-.24	.01
11. Number of smiles	.28	.13	.09	-.31	-.07	-.21	-.03	.13	-,19	-.15		-.07
12. Empathised enjoyment	.78**	.79**	.63**	-.35*	.65**	.58**	.67**	.18	.57**	.51**	.15	

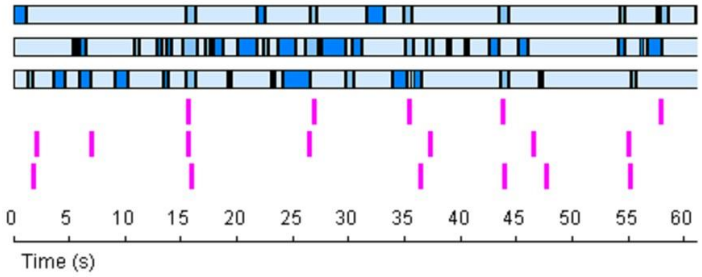
List of figures:

Figure 1. Time-event plot of turn taking and smiling over 60 seconds as computed by The Observer®.

Figure 2. Interaction diagrams showing the number of conversational turns made by each participant (V, W and X) directly following turns made by other group members. Both diagrams show the same group talking about photograph content that supports reminiscing, but in two different equipment configurations.

Figures:

Behavioral Class	Subject
Verbalisation	v
Verbalisation	w
Verbalisation	x
Smiles	v
Smiles	w
Smiles	x



Legend			
Behavior	Color/Pattern	Behavior	Color/Pattern
turn		smile	
utterance			
quiet