Abstract

Relational memory is the ability to flexibly organize and integrate multiple sources of information to produce emergent outcomes. In tests for one type of relational memory - stimulus equivalence - arbitrary stimuli become related in ways not explicitly trained. Little is known however about whether stimulus equivalence-based relational memory ability differentially emerges during offline periods of either sleep or wake. Here, fifty-one, healthy young adults learned a series of interconnected conditional relations involving arbitrary visual images (A-B, A-C, and A-D), and were immediately tested for maintenance of these relations. Following a 12-hour offline period consisting of either sleep or wake, both groups were tested for novel inferences - symmetry (B-A, C-A, and D-A) and equivalence relations (B-C, C-B, C-D, and D-C) - as well as retention of the trained relations. Results from delayed testing, supported by Bayesian statistics, showed that accuracy did not differ between the sleep and wake groups. Potential limitations of this preliminary investigation and directions for future research are discussed.

Keywords: relational memory, stimulus equivalence, sleep, wake, humans.

Long considered to be a hallmark of human cognition, relational memory is the ability to flexibly organize and integrate multiple sources of information (Fantino & Stolarz-Fantino, 2013; Vasconcelos, 2008). Numerous experimental paradigms have been devised to study relational memory, such as transitive inference (TI; Vasconcelos, 2008), associative inference (Preston, Shrager, Dudukovic, & Gabrieli, 2004), and stimulus equivalence (Hayes & Hayes, 1992; Critchfield & Fienup, 2008). In a typical TI task, a series of premise pairs are trained, such as A+B-, B+C-, C+D- and D+E- (where "+" indicates reinforced choices and "-" non-reinforced choices, respectively), before selections from novel combinations of inference pairs (e.g., AE, BD) are tested in the absence of feedback. A period of offline sleep has been shown to facilitate transitive inference in groups matched for premise pair learning, compared with a period of wakefulness (Ellenbogen, Hu, Payne, Titone, & Walker, 2007; Werchan & Gomez, 2013). In the associative inference task (Preston et al., 2004; Zeithamova, Dominick, & Preston, 2012), transitive relations (AC) are tested following overlapping AB and AC training. Lau, Tucker and Fishbein (2010) found a facilitative effect for a non-rapid eye-movement sleep (NREM)-only daytime nap on AC relational memory ability compared to a period of wakefulness.

Unlike the TI and associative inference tasks, stimulus equivalence has been relatively under investigated in the context of relational memory and sleep. This is surprising given that stimulus equivalence has an equally long empirical history as other relational memory tasks (Sidman, 1994), yet, unlike cross-species comparisons of performance on transitive inference and associative inference, there remains no convincing evidence for emergent stimulus equivalence outcomes in nonhumans (Dymond, 2014; Hayes, 1989; Vasconcelos, 2008; but see, Zentall, Wasserman, and Urcuioli, 2014). In a stimulus equivalence task, a series of trained, interconnected relations among physically dissimilar stimuli become related to each other in ways not explicitly trained or instructed (Dymond & Roche, 2013). For example, arbitrary relations involving spoken words, written words and pictures of the corresponding referents illustrate the key properties of stimulus equivalence when symmetry (i.e., trained relations are functionally bidirectional: if A then B yields if B then A) and combined symmetry and transitivity (or equivalence) are present (i.e., trained relations combine in a bidirectional manner: if A then B and if then A then C yields B then C and C then B). For instance, learning that the spoken word “car” is related to or goes with a picture of an actual car (i.e., A-B) and that the spoken word means the same as the written word |car| (i.e., A-C) may lead to untrained, emergent relations. That is, someone given this history will spontaneously match the picture of the car to the written word, and vice versa (i.e., B-C and C-B), and may also utter the spoken word when shown the picture or word and asked, “What is this?” (i.e., B-A and C-A), without further training. Numerous studies over the past four decades have replicated and extended this basic effect (Critchfield, Barnes-Holmes, & Dougher, 2018; Dymond & Roche 2013; Hayes & Hayes, 1992; Sidman, 1994). Relational memory, then, may prove a useful term to describe the emergent outcomes produced by stimulus relations such as stimulus equivalence.

In comparison with findings showing a facilitative role for sleep in other relational memory tasks such as paired associates (Tucker, Tang, Uzoh, Morgan, & Stickgold, 2011), temporal judgment (Drosopoulos, Windae, Wagner, & Born, 2007), semantically related and unrelated word pairs (Payne et al., 2012), or false memory paradigms (Diekelmann, Born, & Wagner, 2010), little is known about the facilitative effects or otherwise of an offline period of sleep on stimulus equivalence-based relational memory ability. To date, evidence does suggest however that once tested, this form of relational memory may be remarkably stable across time. For instance, Saunders, Wachter and Spradlin (1988) investigated this issue in a low-*n* single case design study with participants with developmental disabilities who demonstrated the emergence of equivalence relations during initial testing. Participants were then retested for both the directly trained baseline, as well as derived, relations two- to five-months later. Saunders et al. found that three of the four participants performed at 90% accuracy or better on tests for both baseline and emergent (symmetry and equivalence) relations (Spradlin, Saunders & Saunders, 1992). Similarly, evidence indicates that generalized equivalence relations are retained up to three months after initial testing (Rehfeldt & Hayes, 2000; Rehfeldt & Dymond, 2005). Thus, providing the baseline relations are shown to be intact, equivalence relations may be maintained over time in the absence of intervening learning experiences (see also, Tyndall, Howe and Roche, 2016).

There are however several limitations to this body of research on the long-term stability of stimulus equivalence relations that may have implications for further understanding of the role of sleep in relational memory. First, testing and retesting stimulus equivalence relations after an extended and uncontrolled intervening period does not help to elucidate a role for sleep in how relational knowledge is organised. To unambiguously demonstrate a role for sleep in maintaining stimulus equivalence, it is necessary to conduct an immediate test of the requisite baseline relations in the absence of feedback prior to the intervening period of either sleep or wakefulness and subsequently testing for stimulus equivalence and maintenance of the baseline relations. Second, studies to date on the long-term stability of equivalence have tended to be conducted with developmentally disabled populations and, hence, the determinants of typically developing relational memory effects in sleep remain to be determined. Third, small sample sizes have been used with little or no statistical analysis and an absence of any group comparisons of sleep or wake based intervening period. Fourth, task paradigms have not matched participants for premise or baseline relational learning level at the outset. The stimulus equivalence paradigm therefore confers several advantages as a complementary tool for studying effects of sleep on relational memory. These include controlling for baseline relation learning level by conducting an immediate test of the trained relations prior to the intervening period, the use of arbitrary, physically dissimilar stimuli that participants are unlikely to have encountered before and testing inferential ability across symmetry and equivalence relations requiring complex integration and generalization of relational knowledge.

The present study sought to investigate, for the first time, the effect of an intervening period of sleep on stimulus equivalence based relational memory. Studying sleep and wake effects on stimulus equivalence performance may have implications for future relational memory research and task development. As we have shown, the majority of relational memory paradigms tasks employed to date have either used pre-existing relational categories consisting of familiar stimuli or have not assessed the possible separate or combined effects of the intervening period on both the trained relations and the different types of tested bidirectional relations requiring several intervening steps (e.g., symmetry vs. equivalence relations). Here, we first trained participants to select a series of interconnected baseline relations through trial and error (for the purposes of clarity, labeled here using alpha-numerics: A1-B1, A1-C1, A1-D1, A2-B2, A2-C2, and A2-D2). Participants were then given an immediate test for maintenance of these baseline relations in the absence of feedback. Following a 12-hour offline period consisting of either sleep or wake, both groups of participants returned to the laboratory and were tested for symmetry (B1-A1, C1-A1, D1-A1, B2-A2, C2-A2, and D2-A2) and equivalence relations (B1-C1, C1-B1, C1-D1, D1-C1, B2-C2, C2-B2, C2-D2, and D2-C2), as well as maintenance of the trained relations. On the basis of previous relational memory research (Ellenbogen et al., 2007), we predicted that the sleep group would demonstrate higher percentage accurate responses on the delayed test for stimulus equivalence relations compared with the wake group.

Method

Participants

Fifty-one participants were randomly assigned at the outset to either Sleep (*M* age = 22.5, *SD* = 7.4; 13 women) or Wake (*M* age = 20.4, *SD* = 3.2, 15 women) groups. Sample size was based on, and greater than reported by, previous research (Ellenbogen et al., 2007). Data from one female in the Wake group were excluded from analysis due to failure to reach the relational learning criterion, leaving a final *n* = 25 in each group. No formal exclusion criteria were employed but volunteers were asked to refrain from participating if they had a current diagnosis of either an anxiety or sleep disorder. All participants reported normal sleep-wake histories at the outset. Informed consent was obtained from all participants and the Swansea University Department of Psychology Ethics Committee approved the study.

Apparatus/Materials

Stimuli were twelve abstract geometrical figures, adapted from Vervoort, Vervliet, Bennett, and Baeyens (2014), 4 x 4 cm in size, colored black and presented on white background on a screen positioned at eye level (see Figure 1). The presentation of stimuli and the recording of all responses were controlled by a program written in Visual Basic.NET.

Insert Figure 1 About Here

Participants completed the *Epworth Sleepiness Scale* (ESS; Johns, 1991), the *Pittsburgh Sleep Quality Index* (PSQI; Buysse, Reynolds, Monk, Berman, & Kupfer, 1989), and the *Sleep-Related Behaviour Questionnaire* (SRBQ; Ree & Harvey, 2004). The ESS measures average daytime sleepiness and consists of eight scenarios rated in terms of how likely it is that someone might fall asleep or doze. Answers are scored 0-3, with a total score of 10 or more indicating above average daytime sleepiness. The PSQI measures subjective sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleep medication, and daytime dysfunction over the last month. Answers are scored 0-3, with a total score of 5 or more indicative of “poor” sleeping quality. The SRBQ is a 32-item scale that assesses the use of safety behaviors in insomnia. Items are scored 0-4, with higher scores indicative of greater engagement in safety behaviors in insomnia.

Procedure

On arrival at the Swansea University Sleep Laboratory, participants first completed the ESS, PSQI, and SRBQ. All participants performed an initial training session followed by an immediate test. Participants were first trained on the baseline conditional relations, performed a test session on the relations immediately after learning, and following a 12-hour offline period of either sleep or wake performed a delayed test session that included the baseline relations with symmetry and equivalence relations (Figure 2). Participants in the Wake group were trained at 09.00 (+/- 30 minutes) and completed the delayed test at 21.00 (+/- 30 minutes), while participants in the Sleep group were trained at 21.00 (+/- 30 minutes) and performed the delayed test at 09.00 (+/- 30 minutes) the next day.

Insert Figure 2 About Here

The experiment began when the following instructions appeared onscreen:

> Thank you for agreeing to participate in this study. On the next screen, there will be one item centered at the top of the screen. Your job is to look at the item and then click on it with the computer mouse to display three further items. You should look at the items and then click one of them using the computer mouse. Only one item will be correct: you will receive feedback on your choices. You should try to get as many correct as possible. The more you get right, the quicker you will finish. Please ask the experimenter if you have any questions; otherwise, click “Start” to begin.

During the conditional relational training phase, a delayed matching to sample (DMTS) procedure with feedback was used to present three blocks of 18 trials (54 trials in total; see Figure 3). On each trial, a sample stimulus first appeared centered, at the top of the screen (e.g., A1). After clicking on the sample, it was immediately removed, and three comparison stimuli were simultaneously displayed across the foot of the screen (positions were randomized; B1, B2, B3). Incorrect comparisons were stimuli related to other sample stimuli on other trials (see Figure 3); there was only ever one correct comparison stimulus presented on each trial. Comparisons remained on screen until one was selected by clicking on it with the computer mouse.

Insert Figure 3 About Here

Sample stimuli are labeled here, for the purposes of clarity, A1 and A2, and comparison stimuli are labeled B1, B2, C1, C2, D1, and D2, respectively (participants were never exposed to these labels; see Figure 3). Feedback consisting of the word "Correct" was presented in the center of the screen following correct selections, while the word "Incorrect" was presented following incorrect selections. In the presence of A1, selecting B1, C1, and D1, and in the presence of A2, selecting B2, C2 and D2 was deemed correct (Figure 3). Each trial type (A1-B1, A1-C1, A1-D1, A2-B2, A2-C2, and A2-D2) was presented nine times in a block of 54 trials. Participants were required to achieve a mastery criterion of 92.5% (i.e., 50 out of 54 correct trials) in order to proceed to the next phase. Blocks of trials were repeated until this criterion was met. On meeting the criterion, an immediate test phase was conducted, which consisted of 36 conditional relation trials without any feedback. Each conditional relation trial type (A1-B1, A1-C1, A1-D1, A2-B2, A2-C2, and A2-D2) was presented six times (see Figure 3). Participants then left the lab and were instructed to return at the agreed time; no further task related instructions were given about what to expect when they returned.

On returning 12 hours later, both groups completed the delayed test session (Figure 2), which commenced with the following onscreen instructions:

> Thank you for returning for the final part of this experiment. You will again be presented with one item in the middle of the screen, which you should click on to display three other items. Again, you should select one of the three items by clicking

on it with the computer mouse. You will not receive feedback on your choices. Some of the items may be combined in novel ways; when this happens, use your best guess and what you have learned to make a choice. Click “Start” to begin.

Three types of test trials were presented in one block of 72 trials without feedback: baseline conditional relations (18 trials), symmetry relations (18 trials), and equivalence relations (36 trials). The baseline relations A1-B1, A1-C1, A1-D1, A2-B2, A2-C2, and A2-D2, the symmetry relations B1-A1, C1-A1, D1-A1, B2-A2, C2-A2, and D2-A2, and the equivalence relations B1-C1, B1-D1, B2-C2, B2-D2, C1-B1, D1-B1, C2-B2, D2-B2, C1-D1, C2-D2, D1-C1, and D2-C2 were each presented three times (Figure 3). Trials were presented in a quasi-random order with no more than two consecutive trials of the same type.

Data Analysis

Trials to criterion, percentage correct, and response times were compared with unpaired samples *t*-tests, two-tailed. Repeated measures ANOVA was used to compare response times across different trial types in the combined sample of Sleep and Wake groups. Paired samples *t*-test, two tailed, was used to analyse trends in response times across the combined sample. Alpha was set to .05.

Because of the possibility that our predicted between-group differences may not be supported, we also computed Bayes factors to determine the relative probability of our findings occurring due to either the null hypothesis (H_o) or the alternative hypothesis (H_1) (Rouder, Morey, Speckman, & Province, 2012; Wetzels & Wagenmakers, 2012). Supplementary Bayesian analysis of the behavioural data was conducted (Love et al., 2015) using default priors to estimate the Bayes Factor (BF_{10}), which indicates the relative likelihood of the data occurring under the null hypothesis H_0 and the alternative hypothesis H_1 (Rouder et al., 2012); a BF value greater than 1 indicates a greater likelihood that the data occurred under H_1 than H_0 (Wetzels & Wagenmakers, 2012).

Results

Sleep-related Questionnaires

At the outset, groups did not differ in ESS (Sleep: *M* = 8.08, *SE* = .60; Wake: *M* = 6.92, *SE* = .71), $t(48) = 1.239$, $p = .221$, $BF_{10} = 0.529$, PSQI (Sleep: *M* = 6.04, *SE* = .61; Wake: *M* = 7.6, *SE* = 7.1), $t(48) = -1.649$, $p = .106$, $BF_{10} = 0.851$, or SRBQ scores (Sleep: *M* = 36.0, *SE* = 3.09; Wake: *M* = 41.8, *SE* = 3.31), $t(48) = -1.278$, $p = .207$, $BF_{10} = 0.551$. These findings indicate that levels of daytime sleepiness, sleep quality and sleep related safety behaviour were comparable across the groups.

Performance at Training and Immediate Testing

Groups required a similar number of conditional relation training blocks to meet criterion (Sleep: *M* = 3.44, *SE* = .32; Wake: *M* = 2.84, *SE* = .22), $t(48) = 1.536$, $p = .131$, $BF_{10} = 0.737$. In the immediate test, 17 out of 25 participants in each of the two groups scored 100% correct; however, overall mean accuracy did not differ between groups (Sleep: *M* = 96.8, *SE* = 1.6; Wake: *M* = 97.9, *SE* = .78), $t(48) = -.617$; $p = .540$, $BF_{10} = 0.331$. Similarly, there was no significant difference in time taken to respond (in seconds) to the immediately tested conditional relations (Sleep: *M* = 1.71, *SE* = .11; Wake: *M* = 1.57, *SE* = .05), $t(48) = 1.12$, $p = .26$, $BF_{10} = 0.474$.

Performance at Delayed Testing

Retention of the baseline conditional relations at delayed testing was similar in both groups (Sleep: *M* = 94.6, *SE* = 2.18; Wake: *M* = 93.7, *SE* = 2.65), $t(48) = .257$; $p = .798$, $BF_{10} = 0.290$, with 16 and 17 out of 25 participants in the Sleep and Wake groups, respectively, scoring a 100% accuracy. The mean percentage of correct choices made on symmetry relations test trials (Sleep: *M* = 91.3, *SE* = 2.44; Wake: *M* = 89.9, *SE* = 3.61), $t(48) = .305$; $p = .762$, $BF_{10} = 0.294$, or equivalence relations test trials (Sleep: *M* = 85.3, *SE* = 3.70; Wake: *M* =

82.7, *SE* = 4.17), *t*(48) = .470; *p* = .640, BF_{10} = 0.310, did not differ between Sleep and Wake groups (see Figure 4a).

Insert Figure 4 About Here

Interestingly, the time taken by both Sleep and Wake participants to respond to the delayed test trials varied across the type of tested relations (Figure 4b). A 2 (group) x 3 (relation type: baseline, symmetry and equivalence) repeated measures ANOVA revealed a significant difference in response times, $F(2, 96) = 40.308$, $p < .001$, $\eta p^2 = .453$, BF_{10} = 4.011e, no main effect for group, $F(1, 48) = 1.476$, $p = .287$, $\eta p^2 = .024$, BF_{10} = 0.550, and no interaction of relation type and group, $F(2, 96) = .622$, $p = .539$, $\eta p^2 = .007$, BF_{10} = 2.581e. When Sleep and Wake groups were combined, response times on equivalence relations trials were significantly slower than trials involving both baseline relations, $t(49) = -7.196$, $p <$ .001, BF_{10} = 3.607, and symmetry relations, $t(49) = -7.164$, $p < .001$, BF_{10} = 1.479. These difference in response times across trial types demonstrate experimental control over the delayed testing performance. Overall, however, response times on baseline conditional (Sleep: *M* = 1.99, *SE* = .10; Wake: *M* = 1.76, *SE* = .11), *t*(48) = 1.475; *p* = .147, BF_{10} = 0.685, symmetry (Sleep: *M* = 1.99, *SE* = .14; Wake: *M* = 1.88, *SE* = .12), *t*(48) = .558; *p* = .579, BF_{10} = 0.321), and equivalence relations (Sleep: *M* = 2.58, *SE* = .17; Wake: *M* = 2.33, *SE* = .17), *t*(48) = 1.038, *p* = .304, BF_{10} = 0.439, did not differ between the groups (Figure 4b).

Discussion

Previous studies have shown that an intervening period of sleep facilitates relational memory ability compared to a period of wakefulness. To date, however, no study has investigated stimulus equivalence-based relational memory, in which emergent inferences arise between sets of indirectly related and physically dissimilar stimuli in a complex, bidirectional manner. For the first time, the present study undertook such an investigation and compared intervening 12-hour periods of sleep or wake on the ability to derive arbitrary symmetry and equivalence inferences during a delayed test session.

Groups performed similarly during the immediate test session where baseline relations were trained and tested in the absence of feedback. Accuracy was identical in both groups and confirmed that participants had learned the requisite relations (A-B, A-C and A-D) needed for future derivation or inferential testing. The number of baseline relations training blocks required to reach criterion did not differ between the Sleep and Wake groups, and measured levels of daytime sleepiness, sleep quality and sleep related safety behavior were also comparable across the groups. These results show that groups were well matched at the outset and did not differ in their self-reported sleep related histories and baseline relational learning ability.

During the 12-hour delayed testing of stimulus equivalence relations, specifically of symmetry (B-A, C-A and D-A) and equivalence relations (B-C, C-B, B-D, D-B, C-D, and D-C), groups performed similarly with no evidence of a sleep effect. Accuracy was high across the retested baseline relations and the symmetry and equivalence trials, which were all presented in the absence of feedback. While accuracy was marginally higher in the Sleep group across all trial types, with a linear decline in test accuracy on baseline, symmetry and equivalence trials (Figure 4a), no significant differences were found.

The only between-group differences we observed were in response times during delayed testing. Consistent with previous findings from the stimulus equivalence literature (Barnes-Holmes et al., 2005; Wang & Dymond, 2013), baseline and symmetry relations were solved at similar speeds, with equivalence relations taking longer to solve than both baseline and symmetry. Participants in the Wake group were, on the face of it, marginally faster on all trials in the delayed test, but the absence of a main effect of group or a significant interaction indicates that both groups processed the test tasks at similar speeds.

In contrast to previous findings on relational memory, the present study found, for the first time, no clear benefit for sleep in a stimulus equivalence-based task requiring the integration of multiple sources of information. The high accuracy levels observed on symmetry and equivalence trials may however suggest a potential ceiling effect. Further refinement of our stimulus equivalence task to detect a potential sleep effect (or absence thereof) is needed. However, solving two, four-member (A1-B1-C1-D1 and A2-B2-C2-D2) stimulus equivalence tasks is arguably more demanding than solving five-term series transitive inference tasks (A-B-C-D-E), as the previous research has tended to adopt. Symmetry and equivalence relations require deriving bidirectional links between stimuli separated by a differing number of intervening steps, without feedback. Our findings clearly attest to the fact that participants solved these stimulus equivalence tasks with relative ease on their first test exposure and with no differences in terms of degrees of separation or step-size. In contrast, previous TI studies have reported differences in accuracy as a function of the degree to which item pairs are separated (e.g., 1 step: B-D pair; 2 steps: B-E pair; Ellenbogen et al., 2007; Werchan & Gomez, 2013).

Importantly, failures in both initial acquisition of conditional relations and subsequent derivation of symmetry and equivalence relations have been observed in studies employing similar designs to the present study (Barnes-Holmes et al., 2005; Vervoort et al., 2014). However, these studies undertook immediate testing of stimulus equivalence after the training criterion was achieved. Our findings show that an intervening period of either sleep or wake would not necessarily boost these outcomes.

Stimulus equivalence relations of the size employed in the present study require learning and rehearsing six conditional relations, with responses made in the presence of alternating samples and three comparison stimuli, two of which are incorrect, on every trial. TI tasks, on the other hand, usually present pairs of stimuli with one designated stimulus deemed correct on each trial. The DMTS procedure and observing response made to the sample on every trial in conditional relation training may have facilitated acquisition and long-term retention of the baseline relations and made the task more engaging for participants (Osborne, Heaps, & Phelps-Bowden, 1978; Wilkie & Spetch, 1978). Delayed testing for symmetry and equivalence relations is therefore likely to have been facilitated by the stringent training procedure adopted in the present study that ensured extraneous sources of stimulus control could be ruled out. As such, the training procedures employed compare favourably with previous findings using tasks of larger relational size. For instance, the long-term retention of three, four-member (Rehfeldt & Hayes, 2000) and three, three-member equivalence relations (Rehfeldt & Dymond, 2005) has been shown to be remarkably stable. Further research should however investigate potential sleep effects on relational memory based DMTS tasks with and without observing responses (Hartmann & Warren, 2005).

Moreover, a future study employing the stimulus equivalence paradigm could increase the size of the relational classes from 4 to 5 or 6, increase the number of relations from 2 to 3, and use a training structure other than simultaneous one-to-many (e.g., linear series). These variables have all been shown to influence emergence of stimulus equivalence (e.g., Arntzen, 2012; Ellifsen & Arntzen, 2015; Fields et al., 1997) and should therefore reduce the likelihood of a ceiling effect, allowing for the predicted effects of sleep to be identified.

The present study calculated Bayes Factors (BF) for all analyses to determine the probability of the reported findings occurring due to either the H_o or H_1 (Dienes, 2014; Rouder et al., 2012; Wetzels & Wagenmakers, 2012). Calculating Bayes factors aids interpretation of the relative evidence and may be particularly helpful in between-group comparison studies investigating potential sleep effects on relational memory. Generally, a Bayes Factor greater than 1 indicates a greater likelihood that the data occurred under H_1 than H_0 . In our analyses, only two of the calculated BFs, from the combined group RT analysis, were greater than 1. This indicates that differences in processing speed on baseline, symmetry, and equivalence test trials were due to *H_1* – that solving each relation type requires different relational memory ability. The fact that the Bayes Factor was computed from the combined data set rules out an explanation in terms of the intervening sleep versus wake period. Hence, we may conclude that, sleep does not facilitate a performance speed advantage relative to wake in solving trained or tested trials in a stimulus equivalence task. For the other reported Bayes Factors, our data indicate that findings were due to the *H_o*. That is, sleep and wake groups do not differ in accuracy on trained or tested relations during delayed testing. To our knowledge, this is the first study on potential sleep effects in relational memory to report Bayes Factors. As a means of elucidating null results (Dienes, 2014), Bayesian statistics may also help contribute to initiatives aimed at overcoming so called publication bias (the so-called "file drawer problem") in psychological research (Pashler & Wagenmakers, 2012).

Sleep and wake groups did not differ in terms of age ($p = .255$) but there were two participants in each group who were over 25 years old. These outliers may have influenced group outcomes, particularly as aging is known to affect sleep related consolidation (Harand, et al., 2012) and stimulus equivalence ability when immediately tested (Wilson & Milan, 1995). It would therefore be salutary to conduct a further study that systematically replicates and extends these preliminary findings by including participants from different age groups (Backhaus, Born, Hoeckesfeld, Fokuhl, Hohagen, & Junghanns, 2007).

A final factor worth considering was the experimental design, which compared 12hrs of intervening sleep or wake. Such designs are of course limited because of the potential differences in circadian wake promoting strength, interfering daytime activity, any naps, homeostatic sleep pressure build-up, and caffeine intake, etc. Here, we elected to compare two groups and not include a third group given, for instance, 24hrs of combined sleep/wake intervening period, because previous research on relational memory found no differences between 12hrs and 24hrs of intervening sleep (Ellenbogen et al., 2007). Future studies should however definitively test this assumption in the context of stimulus equivalence by comparing 24hrs and 12hrs sleep groups with an immediate (i.e., 20 min) test group. The results obtained from such a study might permit a clear conclusion to be drawn as to whether or not time and sleep differentially facilitate stimulus equivalence test performance. The present, preliminary findings are limited in this respect by the absence of an immediate testing group. It is noteworthy, however, that the vast majority of studies on stimulus equivalence conduct testing immediately when participants achieve a predetermined training criterion, and that accuracy tends to vary across types of tasks, and number and complexity of stimulus relations tested, and participants studied (Critchfield & Fienup, 2008; Dymond & Rehfeldt, 2001; Green & Saunders, 1998). Thus, while accuracy is likely to be high on an immediate test (reflecting the specific procedures and participant characteristics involved), the relevance for understanding the role of an intervening period of sleep on equivalence based relational memory, and the contribution over that already provided by the Wake and Sleep groups employed in the present study, remains unclear. Furthermore, it will be necessary in any new work to fully assess the role of potential circadian and other variables on relational memory performance. For instance, we did not screen participants in the sleep group for pre-existing sleep problems or clinical disorders that are often treated with psychoactive medication (e.g., depression and ADHD), and nor did we record their sleep quality and duration the night before testing or whether or not they had taken any alcohol or prescription medication.

In conclusion, and notwithstanding the above, the present approach adopted a stimulus equivalence paradigm for the first time to investigate relational memory processes in sleep and wakefulness. It is important to emphasize that the term *relational memory* is intended to refer to a loose collection of empirical phenomena defined by the emergence of untrained stimulus relations. No special status is assumed for the term, and it is hoped that its use here will facilitate greater exchange with stimulus relations researchers from different domains.

Compliance with Ethical Standards

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest: The authors declare no conflicts of interest.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

References

Arntzen, E. (2012). Training and testing parameters in formation of stimulus equivalence: Methodological issues. *European Journal of Behavior Analysis, 13*, 123-135.

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry Research, 28*,193-213.

Backhaus, J., Born, J., Hoeckesfeld, R., Fokuhl, S., Hohagen, F., & Junghanns, K. (2007) Midlife decline in declarative memory consolidation is correlated with a decline in slow wave sleep. *Learning & Memory, 14*, 336-341. doi:10.1101/lm.470507.

Barnes-Holmes, D., Staunton, C., Whelan, R., Barnes-Holmes, Y., Commins, S., Walsh, D., Stewart, I., Smeets, P. M., & Dymond, S. (2005). Derived stimulus relations, semantic priming, and event-related potentials: Testing a behavioural theory of semantic networks. *Journal of the Experimental Analysis of Behavior, 84*, 417-433.

Critchfield, T. S., & Fienup, D. M. (2008). Stimulus equivalence. In S. F. Davis and W. F. Buskist (Eds.), *21st Century Psychology* (pp. 360-372). Sage, Thousand Oaks, CA.

Critchfield, T. S., Barnes-Holmes, D., & Dougher, M. J. (2018). What Sidman did: Historical and contemporary significance of research on derived stimulus relations. *Perspectives on Behavior Science, 41*, 9-32. doi: 10.1007/s40614-018-0154-9

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5: 781.

Diekelmann, S., Born, J., & Wagner, U. (2010). Sleep enhances false memories depending on general memory performance. *Behavioral Brain Research, 208*, 425–429.

Drosopoulos, S., Windae, E., Wagner, U., & Born, J. (2007). Sleep enforces the temporal order in memory. *PLoS ONE*, 4: e376.

Dymond, S. (2014). Meaning is more than associations: Relational operants and the search for derived relations in nonhumans. *Journal of the Experimental Analysis Behavior, 101*, 152-155.

Dymond, S., & Rehfeldt, R. A. (2001). Supplemental measures of derived stimulus relations. *Experimental Analysis of Human Behavior Bulletin*, *19*, 8-12.

Dymond, S., & Roche, B. (2013). *Advances in relational frame theory: Research and application*. Oakland, CA: New Harbinger.

Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D., & Walker, M. P. (2007). Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences of the USA, 104*, 7723–7728.

Ellifsen, C., & Arntzen, E. (2015). Effects of training structure and the passage of time on trained and derived performance. *The Psychological Record, 65*, 1–12.

Fantino, E., & Stolarz-Fantino, S. (2013). The logic and illogic of human reasoning. In G. J. Madden (Ed.), *APA Handbook of Behavior Analysis, Vol.1* (pp. 439-461). American Psychological Association, Washington, DC.

Fields, L., Reeve, K., Rosen, D., Varelas, A., & Adams, B. (1997). Using the simultaneous protocol to study equivalence class formation: the facilitating effects of nodal number and size of previously established equivalence classes. *Journal of the Experimental Analysis of Behavior, 67*, 367-89.

Green, G., & Saunders, R. R. (1998). Stimulus equivalence. In K. A. Lattal and M. Perone (Eds.), *Handbook of research methods in human operant behavior* (pp. 229-262). Springer, New York.

Harand, C., Bertran, F., Doidy, F., Guénolé, F., Desgranges, B., Eustache, F., Rauchs, G. (2012). How aging affects sleep-dependent memory consolidation? *Frontiers in Neurology,* 3:8. doi:10.3389/fneur.2012.00008

Hartmann, M., & Warren, L. H. (2005). Explaining age differences in temporal working memory. *Psychology of Aging, 20*, 645-656.

Hayes, S. C. (1989). Nonhumans have not yet shown stimulus equivalence. *Journal of the Experimental Analysis of Behavior, 51*, 385-392.

Hayes, S. C., & Hayes, L. J. (1992). Verbal relations and the evolution of behavior analysis. *American Psychologist, 47*, 1383-1395.

Johns, M. W. (1991). A new method for measuring daytime sleepiness: The Epworth sleepiness scale. *Sleep, 14*, 540-545.

Lau, H., Tucker, M. A., & Fishbein, W. (2010). Daytime napping: Effects on human direct associative and relational memory. *Neurobiology of Learning and Memory, 93*, 554-560.

Love, J., Selker, R., Marsman, M., *et al*. *JASP* (Version 0.7) [Computer software], 2015.

Osborne, J. G., Heaps, R. S., & Phelps-Bowden, J. (1978). Effect of sample response requirement on matching-to-sample accuracy of exceptional children. *Psychologicsl Reports, 42*, 1267–1276.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the Special Section on Replicability in Psychological Science: A crisis of confidence? *Perspectives on Psychological Science, 7*, 528–530. doi: 10.1177/1745691612465253

Payne, J. D., Tucker, M. A., Ellenbogen, J. M., Wamsley, E. J., Walker, M. P., Schacter, D. L., & Stickgold, R. (2012). Memory for semantically related and unrelated declarative information: The benefit of sleep, the cost of wake. *PLoS ONE* 7(3): e3307.

Preston, A. R., Shrager, Y., Dudukovic, N. M., & Gabrieli, J. D. (2004). Hippocampal contribution to the novel use of relational information in declarative memory. *Hippocampus, 14*,148–152.

Ree, M. J., Harvey, A. G. (2004). Investigating safety behaviours in insomnia: The development of the Sleep-related Behaviours Questionnaire (SRBQ). *Behaviour Change, 21*, 26-36.

Rehfeldt, R. A., & Dymond, S. (2005). The effects of test order and nodal distance on the emergence and maintenance of derived discriminative stimulus functions. *Psychological Record, 55*, 179-196.

Rehfeldt, R. A., & Hayes, L. J. (2000). The long-term retention of generalized equivalence classes. *Psychological Record, 50*, 405-428.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*, 356–374.

Sidman, M. (1994). *Equivalence relations and behavior: A research story*. Boston, MA, Authors' Cooperative.

Spradlin, J. E., Saunders, K. J., & Saunders, R. R. (1992). The stability of equivalence classes. In S. C. Hayes and L. J. Hayes (Eds.), *Understanding verbal relations* (pp. 29-42). Reno, NV: Context Press.

Saunders, R. R., Wachter, J., Spradlin, J. E. (1988). Establishing auditory stimulus control over an eight-member equivalence class via conditional discrimination procedures. *Journal of the Experimental Analysis of Behavior, 49*, 95-115.

Tucker, M. A., Tang, S. X., Uzoh, A., Morgan, A., & Stickgold, R. (2011). To sleep, to strive, or both: How best to optimize memory. *PLoS ONE* 6(7): e21737.

Tyndall, I. T., Howe, B., & Roche, B. T. (2016). Exposure to progressive muscle relaxation leads to enhanced performance on derived relational responding tasks. *Psychological Record, 66*, 213-222.

Vasconcelos, M. (2008). Transitive inference in non-human animals: An empirical and theoretical analysis. *Behavioural Processes, 78*, 313–334.

Vervoort, E., Vervliet, B., Bennett, M., & Baeyens, F. (2014). Generalization of human fear acquisition and extinction within a novel arbitrary stimulus category. *PLoS ONE* 9(5): e96569

Wang, T., & Dymond, S. (2013). Event-related potential correlates of emergent inference in human arbitrary relational learning. *Behavioral Brain Research, 236*, 332-34.

Werchan, D. M., & Gómez, R. L. (2013). An interaction between reinforcement learning and sleep to facilitate transitive inference. *Neurobiology of Learning and Memory, 100*, 70–76.

Wetzels, R., & Wagenmakers, E. J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review, 19*, 1057–1064.

Wilkie, D. M., & Spetch, M. L. (1978). The effect of sample and comparison ratio schedules on delayed matching to sample in the pigeon. *Animal Learning & Behavior, 6*, 273-278.

Wilson, K. M., & Milan, M. A. (1995). Age differences in the formation of equivalence classes. *Journal of Gerontology Series B: Psychological Sciences and Social Sciences, 50*(4), P212-8.

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron, 75*, 168–179.

Zentall, T. R., Wasserman, E. A., & Urcuioli, P. J. (2014). Associative concept learning in animals. *Journal of the Experimental Analysis of Behavior, 101*, 130-151.

Figure Captions

Figure 1. Abstract stimuli used in the current study (Vervoort et al., 2014).

Figure 2. The stimulus equivalence task and experimental paradigm. (A) Twelve abstract stimuli (see Figure 1), eight of which are illustrated here as A1, B1, C1, D1, A2, B2, C2 and D2 (participants were not exposed to these labels), were presented in a delayed matching to sample format. Across conditional relational training trials, in the presence of A1, selecting B1, C1 and D1 was correct, while in the presence of A2, selecting B2, C2, and D2 was correct. During tests of symmetry relations, which were conducted during the delayed test session, it was predicted that participants would select A1 in the presence of B1, C1, and D1, and select A2 in the presence of B2, C2 and D2, without further feedback. During tests of equivalence relations, it was predicted that participants would select B1 given C1 and D1, C1 given B1 and D1, D1 given B1 and C1, B2 given C2 and D2, C2 given B2 and D2, and D2 given B2 and C2, without further feedback. (B) To determine the effects of sleep on the formation of stimulus equivalence relations, all participants first learned conditional relations. Immediately after learning had occurred, all participants were tested on the conditional relations to determine the extent of learning without feedback. Then, following an offline delay interval consisting of either 12 hours sleep or 12 hours wakefulness, participants were again tested on the conditional relations and were also tested for their ability to reverse and combine stimulus relations by testing for symmetry and equivalence relations.

Figure 3. Schematic overview of the trained and tested relations. The upper panel shows trials used in the first session (conditional relations training and testing) and the lower panel shows trials used in the second session (conditional relations testing, symmetry relations testing and equivalence relations testing). The first stimulus on the left of each array represents the

sample stimulus; the other three stimuli are the comparison stimuli for that particular trial. The correct comparison stimulus is indicated in **bold**.

Figure 4. Delayed test performance. (a) Baseline conditional relations (averaged across all conditional relations A-B, A-C, and A-D), symmetry relations (averaged across all novel test relations B-A, C-A and D-A), and equivalence relations (averaged across all novel test relations B-C, C-B, C-D, D-C, B-D, and D-B) performance at the delayed test session for both groups. (b) Response times on baseline conditional relations, symmetry relations and equivalence relations trials the delayed test session for both groups. Error bars represent standard errors.