Psychiatry Research 265 (2018) 200–207

Contents lists available at ScienceDirect

Psychiatry Research

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

Towards a reliable repeated-measures beads task for assessing the jumping to conclusions bias

Benjamin F. McLean[a,b], Julie K. Mattiske[a,*], Ryan P. Balzan[a,b]

[a] College of Education, Psychology and Social Work, Flinders University, Adelaide, Australia
[b] ARC Centre of Excellence in Cognition and its Disorders, Macquarie University, Sydney, Australia

ARTICLE INFO

Keywords:
Beads task
Jumping to conclusions
Repeated measures
Repeatability
Reliability
Psychosis

ABSTRACT

The jumping to conclusions bias (JTC), in which some people gather less information than others before making a decision, has been linked to delusions in psychosis. JTC is usually identified via the beads task, in which a sequence of beads (the "target" sequence) is used to measure the amount of evidence participants require before making a decision. Yet, despite its common use, the reliability of the task has never been properly investigated. We investigated its reliability, and tested an alternate version which used distractor sequences to obfuscate the target sequence. Healthy participants ($N = 212$) were randomised into two groups. One group completed ten trials using the target sequence, while the other completed ten trials of the target sequence and three distractor sequences. Our data indicated the standard task may not be reliable over repeated measures, but that by including distractor sequences, the task becomes more believable, repeatable, and reliable. Additionally, excluding first-trial data (a "silent" practice trial) also improves repeatability. These improvements to the task are relevant to single trial studies, and will be especially useful to repeated-measures longitudinal, experimental, and treatment studies. Such repeated-measures studies are important for investigating the causal link between JTC and delusions.

1. Introduction

The beads task (Huq et al., 1988) is used to measure the *jumping to conclusions* bias (JTC), in which some people gather less information than others might do before making a decision. Typically, participants are presented with two jars containing beads of two colours in equal but opposite ratios. One jar is chosen at random, and beads are drawn one at a time (purportedly at random, but actually in a predetermined sequence) from the jar. At each draw, participants either make a decision as to which jar beads were being drawn from, or request to see more beads. The amount of data gathered is most often measured as the number of beads drawn before a decision is made ("draws to decision" – DTD), though other criteria are possible, such as Liberal Acceptance (in which participants make a decision at a low level of confidence; Moritz et al., 2007). The beads task has been key in the large body of research linking JTC with delusions in psychosis (Dudley et al., 2016; McLean et al., 2017; Ross et al., 2015). However, despite the sizeable beads-task literature (a search of PubMed Central for the terms "*JTC*", "*jump(ing) to conclusions*", or "*beads task*" returned 1390 full-text journal hits), the reliability of the beads task has never been seriously investigated.

In the large majority of beads-task studies only a single trial of the beads task is presented – and when more than one trial is presented, the same sequence of beads is generally used repeatedly. Yet it is unclear whether DTD captured over single or multiple trials is a genuine reflection of a participant's true DTD. Two studies have found acceptable test-retest reliability over two and three trials (Lincoln et al., 2010; Moritz et al., 2015). However, and concerningly, two recent meta-analyses have shown that responses on the beads task vary with the number of trials provided (Dudley et al., 2016; Ross et al., 2015). There is also some evidence that between-group differences in JTC may only be observable on a single trial, and may disappear over multiple trials (Krug et al., 2014; Rausch et al., 2014). That the beads task may not be consistent across repeated measures signals possible problems with the psychometric properties of the task both in single-trial studies, and in longitudinal, treatment, and experimental studies that employ repeated administrations. Thus, our first objective was to assess consistency over multiple trials.

Changes in response to the beads-task over repeated measures might occur for a variety of reasons. Perhaps participants' responses change when they discover the predetermined nature of the bead sequence with which they are presented. Participants repeatedly exposed to the

* Corresponding author at: College of Education, Psychology and Social Work, Flinders University, Adelaide, Australia.
E-mail addresses: ben.mclean@flinders.edu.au (B.F. McLean), julie.mattiske@flinders.edu.au (J.K. Mattiske).

https://doi.org/10.1016/j.psychres.2018.04.043
Received 8 November 2017; Received in revised form 29 March 2018; Accepted 12 April 2018
Available online 13 April 2018
0165-1781/

AAABAAAABA sequence in common use (the "target" sequence) may soon observe that the first few beads are always of one colour, and that the first colour is always the majority colour, and begin to make earlier decisions than previously. Bead colours are usually changed from trial to trial to try and prevent this (see for example Ross et al., 2011; Waller et al., 2011), however simply changing bead colours may not effectively hide the single repeating target sequence.

Alternatively, perhaps early beads-task responses are influenced by the range of unknowns which participants undoubtedly face when completing this task for the first time. Supporting this idea, evidence from the broader area of cognitive psychology has found that practice effects over the first two trials of a range of cognitive assessment tools can produce significant changes, while responses over later trials are more consistent (Collie et al., 2003). If this is the case, it is of particular concern that the beads task is usually only administered once, and with no practice trial (Dudley et al., 2016).

Other measures of JTC have the potential to avoid the problem of repeating identical sequences, such as the box task (Balzan et al., 2017; Moritz et al., 2017), however the task is relatively new and its validity has not yet been established. The non-serial data-gathering paradigm of van der Leer et al. (2017), in which participants request the full number of fish they wish to see drawn from a lake in one go, could also avoid the issues with repeating sequences. Again however, this task is a significant divergence from the standard beads task, and requires further validation. In the meantime, the beads task is in many ways the standard method of identifying JTC, and its limitations over repeated measures needs to be established and improved.

To address the potential problem of participants recognising the key characteristics of the beads-task target sequence over repeated measures (which may result in changes in response), we proposed interleaving this sequence between distractor sequences. We expected distractor sequences to reduce the salience of the target sequence and, therefore, participants' recognition of its main features. We tested this solution using a two-group (target-sequence-only and distractor-sequences groups) repeated-measures design, in which each group completed 10 trials of their respective beads task. In addition, to assess the efficacy of a practice trial to address the potential problem of participants responding less reliably on initial trials, we compared the consistency of two consecutive trials conducted with and without a practice trial.

In order to assess the efficacy of distractor sequences in disguising the fact that the beads sequences were non-random, we recorded participants' self-reported belief that they were random. To assess the consistency over repeated trials of the target-sequence-only and distractor-sequences tasks, we measured the repeatability and reliability of the beads task over repeated measures. Here repeatability means that a measure returns stable or consistent values over trials under conditions in which a construct's true level should not change (e.g., measurements taken close in time, under identical conditions, with no intervention between measurements). This was operationalised as the mean squared error (MSE) over repeated measurements, reflecting intra-individual stability. Reliability on the other hand is the ability to discriminate between levels of a variable in the presence of noise (Portney and Watkins, 2015). This was operationalised as the Intraclass Correlation Coefficient (ICC(1,1), Shrout and Fleiss, 1979), which is the proportion of total variance explained by the true variance due to genuine differences between people.

The impact of having a reliable repeated-measures beads task would be significant. Such a task is needed for longitudinal, experimental, and treatment studies, and whenever aggregate measures are to be calculated for increased accuracy and precision.

2. Methods

2.1. Participants

240 participants on the Prolific online crowdsourcing platform completed our experiment. Participants were paid £2.20 for participation.

2.2. Beads task

Every participant completed ten trials of either the target-sequence-only or distractor-sequences beads task. The target-sequence-only group completed ten trials using only the target sequence (AAABAAAABA), while the distractor-sequences group completed ten trials using the target sequence plus three additional distractor sequences (i.e., the target sequence was presented ten times to this group also, but along with three distractor sequences each time it was presented). Comparing the two groups on an equal number of target sequences necessitated the distractor-sequences group completing 40 beads-task sequences compared with the 10 completed by the target-sequence-only group. The distractor sequences included one sequence for which the first bead colour was the minority colour, as this would occur from time to time under truly random conditions. The distractor sequences were AABAAABAAA, BAAAABAAAA, and AAAABAABAA, with the target sequence being presented between the first and second distractor sequences. New trials were identified via on-screen text, and each trial used a unique pair of bead colours.

For each sequence, the identity of the majority colour "A" and its association with the left or right-hand jar were pseudo-randomised. Participants were initially presented with a single bead, along with the query "Would you like to make a decision regarding which jar beads are being drawn from?". Participants could either select "no, I would like to see another bead", or "yes, I have made a decision". Whenever participants requested another bead, the next bead in the sequence (to a maximum of ten beads) was displayed on screen, along with any previous beads to ensure responses were not affected by memory capacity (Freeman et al., 2014). When participants elected to make a decision, the jar they chose along with the DTD were recorded, and the sequence was stopped.[1] If no jar was chosen after the 10th bead, a DTD of 11 was recorded and the participant was progressed to the next sequence. This continued until participants completed ten trials of one sequence (target-sequence-only group) or ten trials of four sequences (distractor-sequences group).

2.3. Procedure

Participants accessed the online experiment via their own computer device, and were randomised to either the target-sequence-only condition or the distractor-sequences condition. After informed consent was established and basic demographic data were collected, participants were presented with detailed instructions explaining the beads task appropriate to their group, and a comprehension check consisting of a two-item test of their understanding of the task (see Experiment 10, Crump et al., 2013). If a participant failed the comprehension check it was repeated until a correct answer was recorded. Participants then completed ten beads-task trials.

Following completion of the beads task, the Peters Delusions Inventory (PDI, Peters et al., 2004) was administered (PDI data are not reported in this study), along with an attention check part-way through the 21 items. Participants then completed Likert scales (from 0 = definite disbelief to 4 = definite belief) indicating their belief in the non-swapping of beads-task jars mid-trial and in the randomness of

[1] Participants were not asked for their degree of confidence once they had made their decision.

Table 1
Demographics and descriptive statistics.

	n	Gender	Age	Mean trial 2–10 DTD			Mean trial 2–10 trial-pair MSE		
			M [95% CI]	*Mdn* [IQR]	Skew. *z*	Kurt. *z*	*Mdn* [IQR]	Skew. *z*	Kurt. *z*
Target-sequence-only group	106	46 female, 60 male	32.3 [29.9, 34.9]	3.11 [2.08, 4.89]	3.55**	1.58	0.23 [0.00, 0.94]	15.30**	1.58
Distractor-sequence group	106	48 female, 58 male	31.4 [29.0, 33.8]	2.78 [2.00, 3.44]	7.30**	9.44**	0.16 [0.00, 0.40]	13.41**	30.56**

** significant at the 0.01 level (i.e., $z > 2.330$).

sequences. Those who did not express definite belief in either aspect were asked to estimate the trial on which they first questioned the task, and how this affected their responses (whether causing them to increase, decrease, or maintain DTD) (not reported in this study). Finally, an instructional manipulation check (IMC, Oppenheimer et al., 2009) was provided to test whether participants were, at this stage of the experiment, reading instructions in their entirety. As this IMC involved a "trick" question and could prompt systematic thinking (Hauser and Schwarz, 2015) it was presented last, so as not to affect participants' responses on other measures.

2.4. Statistical analyses

Each group's relative belief in the randomness of sequences was established by an independent-samples *t*-test, and its association with repeatability was assessed by a Spearman's ρ correlation. We assessed the impact of a practice trial by comparing the within-subjects' repeatability of trials 1 and 2 (equivalent to a two-trial design with no practice trial) against the repeatability of trials 2 and 3 (equivalent to a two-trial design with a "silent" practice trial). We used non-parametric tests (see below) to check for both systematic (change in DTD) and random (MSE) changes in response over trials to establish whether a practice trial provided greater repeatability. We then tested the repeatability of each task over multiple trials by comparing MSE between groups. Due to the non-normal distribution of the DTD and MSE data (see below), all DTD and MSE comparisons were made using the non-parametric Wilcoxon Signed-ranks test (related-samples) or Mann–Whitney *U* test (independent samples). Non-parametric effect sizes were expressed as rank-biserial correlations (Mann–Whitney *U* tests) or matched-pairs rank-biserial correlations (Wilcoxon Signed-ranks tests) calculated according to the formulas of Glass (1965) and Kerby (2014) respectively. The rank-biserial correlation (*r*) is an easily-interpreted 'common language effect size' (McGraw and Wong, 1992) which is calculated as $P_{for}/P_{total} - P_{against}/P_{total}$, i.e., the proportion of ranked pairs favouring a hypothesis minus the proportion of ranked pairs favouring its antithesis (Kerby, 2014).[2] This results in a value within the range [1, −1] expressing the strength of favourable evidence.

To assess reliability, we computed the Intraclass Correlation Coefficient (ICC(1,1), Shrout and Fleiss, 1979) of DTD over increasing ranges of trials, beginning at trial 2. This provides the average reliability of any one trial within the given trial range. We calculated ICC(1,1) by the ANOVA estimation method (Donner and Wells, 1986). However, it was necessary to hold the estimated true variance constant, because being a ratio of true variance to total variance, ICC values are not comparable if the estimated true variance is permitted to change over trial ranges or between groups (Cohen and Doveh, 2005; Hanges and Lyon, 2005). We estimated true variance from the mean square between and mean square error of the distractor-sequences group over trials 2 to 3[3] (Shrout and Fleiss, 1979).

[2] Tied pairs were counted as 0.5 for and 0.5 against.
[3] Chosen as our results indicate they are the least affected by error.

3. Results

3.1. Data screening

Nineteen participants were excluded for either failing the attention check ($n = 5$), and/or the IMC more than twice ($n = 9$), and/or for answering more than 30% of target trials incorrectly ($n = 7$). We then inspected box-plots of the DTD and MSE ten-trial averages, and found a small number of extreme responses more than four inter-quartile ranges from the median. These extreme responses (4 MSE responses from the target-sequences group, and 2 MSE and 4 DTD responses from the distractor-sequences group) may reflect responding that was influenced by miscomprehension or a major individual difference on an unmeasured third variable rather than true construct levels, and were excluded from the data. Final participants were 94 females and 118 males ranging between 18 and 64 years old. Groups were well-balanced demographically (see Table 1). The distribution of mean trial-2-to-10 DTD and of mean trial-pair MSE over trials 2-to-10 were mostly significantly positively skewed and leptokurtic (Kim, 2013; Tabachnick and Fidell, 2007), so nonparametric tests were used for DTD and MSE comparisons. Both groups indicated probable to certain belief that jars were not swapped during trials (target-sequence-only group: $M = 3.60$, $SD = 0.84$, 95% CI [3.44, 3.76]; distractor-sequences group $M = 3.40$, $SD = 1.08$, 95% CI [3.19, 3.60]), and it appears unlikely miscomprehension of the task (Balzan et al., 2012a,2012b) was a significant issue. Trial-to-trial DTD is shown on Fig. 1.

3.2. Belief in the task

The primary purpose of introducing distractor trials to the beads task was to hide the non-random nature of the target sequence. Belief in the randomness of sequences was significantly higher for the distractor-sequences group ($M = 2.32$, $SD = 1.45$, 95% CI [2.04, 2.60], indicating ambivalent to probable belief) than for the target-sequence-only group ($M = 1.68$, $SD = 1.44$, 95% CI [1.40, 1.96], indicating ambivalent to probable *disbelief*), $t(210) = -3.23$, $p = 0.001$, Hedges' $g = 0.44$.

3.3. Practice trial

The effect of a practice trial was assessed by comparing the repeatability of a practice-task (trials 2–3, discarding trial 1 data) and no-practice-task (trials 1–2) subset of trials. We considered both within-subjects' (MSE) and systematic data (group-level change in DTD, Fig. 1), as it is possible for a stable group-level response to hide significant but random within-subjects' changes, or for non-significant but systematic within-subjects' changes to result in significant group-level changes. For the target-sequence-only group, a Wilcoxon signed-ranks test showed that change in DTD was greater over the no-practice-task trials than over the practice-task trials, indicating that group DTD was more consistent over trials when trial 1 was treated as a practice trial (Table 2). However, within-subjects' MSE over the practice-task trials was not significantly different from MSE over the no-practice task trials, indicating that a practice trial did not improve within-subjects' repeatability in this group. For the distractor-sequences group, there was no significant difference in change in DTD over the no-practice-task

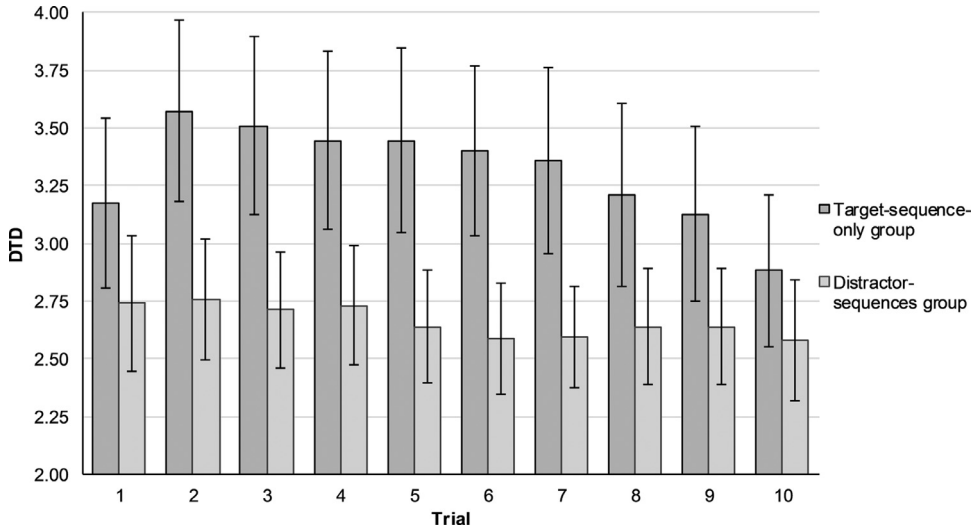

Fig. 1. Mean DTD at each trial, with 95% confidence intervals.

Table 2
Descriptive and test statistics for repeatability with and without a practice trial.

			Mdn [IQR]	*M [95% CI]*	*Z*	*p*	*R*
Change in DTD	Target-sequence-only group	With practice trial	0.00 [0.00, 0.00]	−0.07 [−0.28, 0.14]	−2.54	0.011	−0.160[a]
		Without practice trial	0.00 [0.00, 1.00]	0.42 [0.20, 0.63]			
	Distractor-sequences group	With practice trial	0.00 [0.00, 0.00]	−0.05 [−0.17, 0.08]	−0.21	0.831	−0.028[b]
		Without practice trial	0.00 [0.00, 0.00]	0.02 [−0.19, 0.22]			
MSE	Target-sequence-only group	With practice trial	0.00 [0.00, 0.25]	0.34 [0.22, 0.46]	−1.15	0.251	−0.085[c]
		Without practice trial	0.00 [0.00, 0.25]	0.29 [0.18, 0.41]			
	Distractor-sequences group	With practice trial	0.00 [0.00, 0.25]	0.28 [0.17, 0.38]	−3.45	0.001	−0.189[d]
		Without practice trial	0.00 [0.00, 0.25]	0.11 [0.06, 0.15]			

Comparisons are
[a] Group-level changes – target sequence only group.
[b] Group-level changes – distractor-sequences group.
[c] Within-subjects' changes – target sequence only group.
[d] Within-subjects' changes – distractor-sequences group.

trials compared with the practice-task trials, indicating that a practice trial did not improve group-level repeatability. However, within-subjects' MSE over the practice-task trials was less than MSE over the no-practice task trials, indicating that a practice trial improved within-subjects' repeatability in this group. Together these results showed that using a practice trial increased group-level (target-sequence-only task) and within-subjects' (distractor-sequences task) repeatability, and trial-1 data were therefore excluded from subsequent analyses.

3.4. Repeatability comparison

To test whether distractor sequences improved repeatability we calculated the MSE of each group over ranges of trials commencing at trial 2 (Fig. 2). Mann–Whitney *U* tests showed that the MSE of the distractor-sequences group was significantly less than that of the target-sequences-only group for all trials ($U = 4457.5–4622.5$, $Z = -2.20–3.09$, $p = 0.002–0.028$, $r = 0.170–0.207$), indicating that the distractor-sequences group made smaller changes in DTD response over all ranges of repeated trials than did the target-sequences-only group. Hence using distractor sequences produced a more repeatable beads task. As we had predicted that increased belief in the randomness of beads sequences would lead to increased repeatability of responses, we tested whether increased repeatability was associated with increased belief. A Spearman's rho correlation indicated higher belief was significantly associated with lower MSE in the target-sequence-only group, $\rho = -0.387$, $p < 0.001$. No association was evident in the distractor-sequences group, $\rho = -0.004$, $p = 0.969$, however the majority of participants in this group may have had sufficient belief in the randomness of sequences to respond authentically, and with high repeatability (68% expressed ambivalent to certain belief in the randomness of sequences).

3.5. Reliability comparison

To compare groups on their reliability (the ability to discriminate between levels of a variable in the presence of noise) we computed the Intraclass Correlation Coefficient (the ratio of true to total variance) of DTD over various ranges of trials, beginning at trial 2 (Fig. 3). The reliability of responses to the target sequence in the distractor-sequences task was, for all trial ranges, greater than that of the target-sequence-only task. For the distractor-sequences task, true between-subjects' differences explained 89% of the total variance in scores over trials 2–3, and over all ten trials explained 84% of the variance in scores on average. In contrast, for the target-sequence-only task, true between-subjects' differences explained only 74% of the variance in scores over trials 2–3,[4] and over all ten trials just 62% of the variance in scores on average – poor reliability for a standardized task, repeated without delay under identical conditions.

[4] 71% when no practice trial was used (as is common in the field).

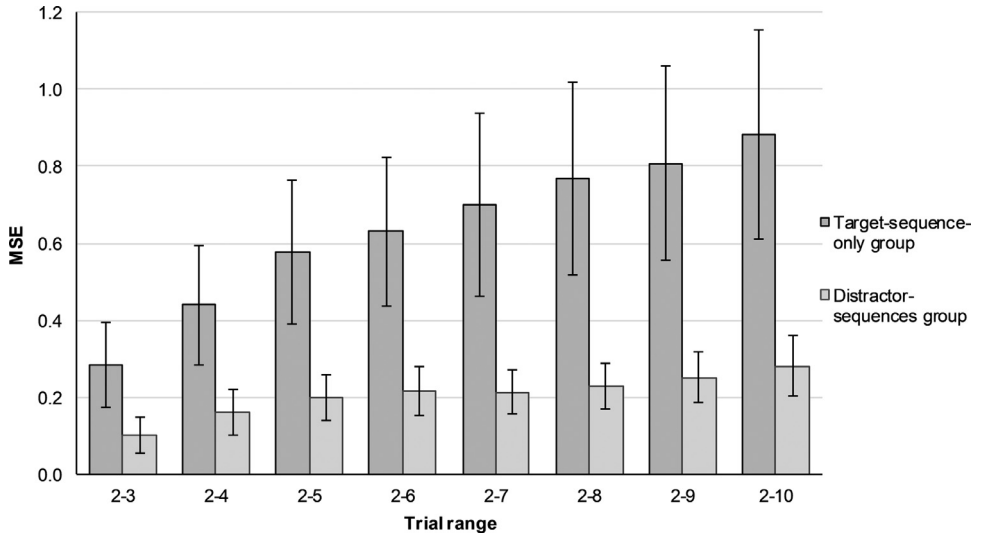

Fig. 2. Mean MSE over increasing ranges of trials, with 95% confidence intervals.

4. Discussion

The beads task is commonly implemented as a single beads sequence, whether in single-trial or pre-/post-intervention designs, but may suffer from reliability and validity problems. We assessed the stability of DTD over multiple repetitions of a single beads sequence, tested the effect of distractor sequences on the repeatability and reliability of the beads task, and explored whether distractor sequences increased participants' belief in the authenticity of the task. We also assessed the impact of using a practice trial.

4.1. Repeatability

The original beads task, consisting as it does of only one non-random sequence of beads, is likely to be ill-suited to repeated measures. Participants exposed to this sequence over multiple trials, especially in quick succession, can hardly be expected *not* to realise the sequence is pre-determined. Our expectation was that this realisation would lead to changes in response to the task, and poor repeatability, over repeated measures – but that increasing believability by using distractor trials might be one way to increase repeatability. The MSE (representing within-subjects' changes over trials) of the target-sequence-only group increased monotonically with every repetition of the task (Fig. 2), indicating poor repeatability, as we predicted. Meanwhile, as expected, using distractor sequences resulted in better repeatability, no matter how many trials participants completed (up to the nine trials tested).

Repeatability is a necessary quality for a repeated measure. Improving the repeatability of the beads task will make it more useful as a measure of data-gathering and JTC over the repeated measures necessary to conduct longitudinal, experimental, and treatment studies. Such studies are becoming increasingly important in the field of delusions research, where JTC has been linked with greater delusional severity in people with psychosis (Dudley et al., 2016; McLean et al., 2017; Ross et al., 2015). To date this association has been observed mainly in observational and correlational studies, but longitudinal, experimental, and treatment studies are of increasing interest for inferring causality (e.g. Moritz et al., 2013; Pos et al., 2017; Rocha and Queiros, 2013; Ross et al., 2011), and the importance of having a repeatable beads task is correspondingly greater. Our data indicate that the distractor-sequences beads task is safely repeatable over at least nine trials with little deterioration. These trials were conducted in immediate succession, and represent a conservative test of repeatability; adequate repeatability over longer time-periods is all-the-more likely to

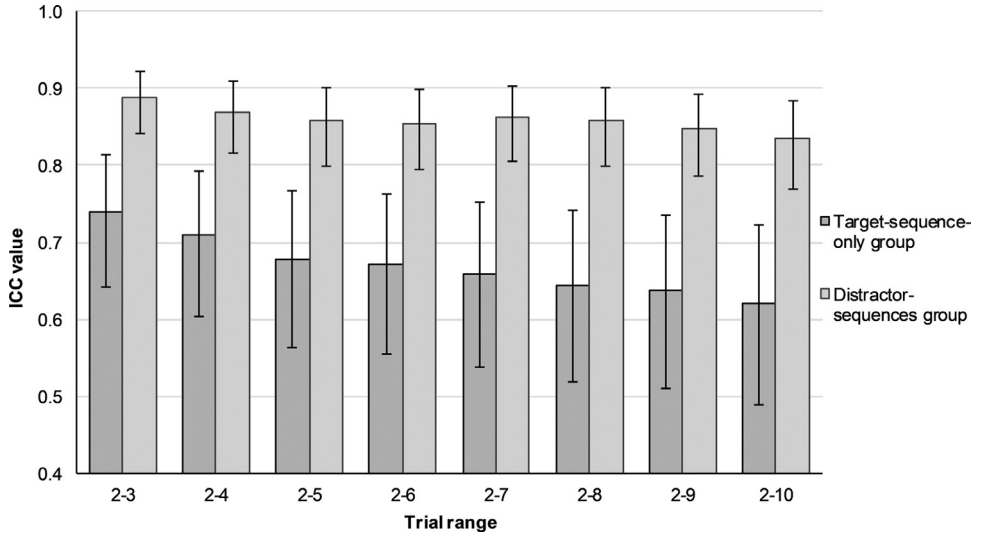

Fig. 3. ICC(1,1) reliability over increasing ranges of trials, with 95% confidence intervals.

be achievable.

4.2. Reliability

The reliability of a measure places limits on its ability to differentiate between levels of a construct in the presence of noise, and hence is critical to its usefulness. The distractor-sequences beads task was considerably more reliable than the target-sequence-only task (Fig. 3). In particular, the reliability of the target-sequence-only task was low for such a standardised task, repeated without delay under identical conditions. Hence, the distractor-sequences beads task is significantly better able to reliably separate people on their tendency to gather data and on the JTC construct.

The greater reliability and repeatability of the distractor-sequences beads task supports its use in the repeated-measures longitudinal, experimental, and treatment studies so necessary to the delusions and cognitive bias field at present (Rausch et al., 2016). It will also maximise the precision benefits available by averaging multiple DTD measurements (Cumming, 2012, p. 377), and better allow for exploring subtle effects such as the uncertain relationship between DTD and delusion-proneness in healthy samples (see for example the small effect size and variety of study results in the meta-analysis by Ross et al., 2015). Future research should establish the reliability and repeatability of the distractor-sequences beads task over greater time intervals however, as while some experimental designs may call for rapid repetitions, many experimental designs, including treatment studies, are conducted over extended time intervals (i.e., weeks or months).

While the high repeatability and reliability of the distractor-sequences task are encouraging, because the use of distractor sequences quadruples the number of sequences participants face compared to the standard beads task, the consequences of fatigue or disengagement with the task must also be considered. Fatigue may have the capacity to change what the task is measuring, away from the intended variable of the tendency to gather information – although that the distractor-sequences group did not significantly change DTD or have large MSE over 10 trials would tend to indicate against this. Nevertheless, the task has not been tested in clinical samples, and it is quite possible that participants with schizophrenia for example, with attendant cognitive deficits (Freeman et al., 2014; Ochoa et al., 2014), may find the extra trials more fatiguing. Thus, the distractor-sequences task would benefit from further validation in both clinical and non-clinical samples.

4.3. Task believability

Our purpose in introducing distractor trials to the beads task was to hide the non-random nature of the target sequence, thus increasing the believability of the task, and so increasing task repeatability and reliability. The results support our argument that distractor sequences would increase task believability - the distractor-sequences group demonstrated a stronger belief in the randomness of beads sequences than the target-sequence-only group with a near medium effect size.

Increased believability was associated with greater repeatability of responses in the target-sequences-only group, and we speculate that greater belief in the task was one factor which lead to greater repeatability and reliability. Belief in the task also impacts upon task validity. For the beads task to be valid as a measure of data gathering, participants must believe there is a genuine unknown (which jar beads are being drawn from), which they can solve by gathering genuine data (by requesting beads). When this is not the case, responses may be less valid. For example, perhaps participants choose more beads once they question the authenticity of the task to see if the sequence is random or not; others may choose less beads in the knowledge more beads will not provide additional information on jars. Such changes in response may be evident in the higher within-subjects' MSE of the target-sequence-only group. Hence uncertain belief may lead to inauthentic responses and low task validity, along with poor repeatability and decreased reliability.

It is possible that when the beads task is administered only once using a single sequence, participants might have high belief in the randomness of the single sequence to which they are exposed. Nevertheless, meta-analyses (Dudley et al., 2016; Ross et al., 2015) have shown DTD tends to change with the number of beads-task trials, indicating that DTD results from a single trial may not be accurate, however believable the single trial. In contrast, the distractor-sequences beads task improves repeatability and reliability of the task over repeated measures, while maintaining a high level of task believability.

Our believability data have limitations however. Simply asking the question "Do you think beads were presented to you in random sequences?" may have implied to participants that bead sequences were not random, and hindsight bias (Guilbault et al., 2004) or desirability factors (Ganster et al., 1983; Podsakoff et al., 2003) may have lead some participants to underestimate or understate the degree to which they believed that sequences were random. Future studies might address this limitation by measuring belief in the randomness of sequences by methods less subject to hindsight bias and desirability effects, for example by indirect questioning (Fisher, 1993; Jo et al., 1997). Nevertheless, self-reported belief in the randomness of sequences was higher in the distractor-sequences group, counter to social desirability.

4.4. Practice trial

In their meta-analysis, Dudley et al. (2016) considered the use of practice or multiple trials a key factor in the validity of studies which measured JTC using the beads or equivalent tasks, and our results certainly support their assessment. We found that participants in both groups changed their responses significantly more over trials 1-to-2 than over trials 2-to-3, indicating that a practice trial can improve the repeatability of responses to the beads task.

While some beads-task studies have used a practice trial (e.g. Brankovic and Paunovic, 1999; Dudley et al., 2011), the large majority have not (Dudley et al., 2016). The lack of a practice trial and hence decreased repeatability may have weakened results in some cases. For example Balzan et al. (2012b) noted in their repeated-measures study without a true practice trial[5] that practice effects may have contributed to observed changes, a concern which our study would support. Alternatively, implementing a practice trial may attenuate between-group differences in JTC, and the impact of a practice trial on group JTC differences requires further research.

It is possible a practice trial could help correct aspects of miscomprehension regarding the beads task, such as the potential for participants to mistakenly think jars are swapping whenever bead colours swap mid-trial (Balzan et al., 2012a,2012b). However, our task instructions were extensive on all key aspects, and participants reported believing jars did not swap mid-trial. The benefit of a practice trial may instead be mostly in familiarising participants with the task, thereby reducing the impact of any factors which could differentially affect participant responses as their familiarity and comfort with the task changes. Future research should also test the effect of providing a practice trial on the graded-estimates version of the beads task, in which miscomprehension was first observed (Balzan et al., 2012a,2012b).

It must be noted that our practice trial was not identified as such to participants. Instead we excluded data from the first trial in a series, an approach that was used by Brankovic and Paunovic (1999) to ensure comprehension and to demonstrate the randomness of the beads sequence to participants. Hence the practice-trial results strictly apply only to situations in which a "silent" practice trial of an identical nature to other trials is provided but not identified as such to participants.

[5] Participants were shown a demonstration trial.

Arguably, participants may perceive and treat explicitly-declared practice trials differently from silent practice trials they presume are being scored, and future studies should test whether a silent practice trial or an explicitly declared practice trial provides the better reliability. The effect of providing feedback following a practice trial could also be investigated.

4.5. General conclusions

To our knowledge, no beads-task studies have attempted to mask the obvious, non-random nature of the target sequence other than by simple colour or ratio changes. By surrounding the target sequence with three distractor sequences, we have developed a beads task that is demonstrably more believable, repeatable, and reliable. We propose that with further validation, the distractor-sequences beads task, accompanied by a "silent" practice trial, be used in place of the traditional beads task – both whenever a single measure of the data gathering construct is needed, and especially when repeated or aggregate measures are required.

Acknowledgements

We thank R. Dudley for supplying background data from the Dudley et al. (2016) meta-analysis, and P. Skuza for providing statistical consultation regarding repeatability, reliability, and the ICC.

Financial support

None.

Conflict of interest

The authors declare no conflict of interest.

References

Balzan, R., Delfabbro, P., Galletly, C., 2012a. Delusion-proneness or miscomprehension? A re-examination of the jumping-to-conclusions bias. Aust. J. Psychol. 64 (2), 100–107. http://dx.doi.org/10.1111/j.1742-9536.2011.00032.x.

Balzan, R.P., Delfabbro, P.H., Galletly, C.A., Woodward, T.S., 2012b. Over-adjustment or miscomprehension? A re-examination of the jumping to conclusions bias. Aust. N. Z. J. Psychiatry 46 (6), 532–540. http://dx.doi.org/10.1177/0004867411435291.

Balzan, R.P., Ephraums, R., Delfabbro, P., Andreou, C., 2017. Beads task vs. box task: the specificity of the jumping to conclusions bias. J. Behav. Ther. Exp. Psychiatry 56, 42–50. http://dx.doi.org/10.1016/j.jbtep.2016.07.017.

Brankovic, S.B., Paunovic, V.R., 1999. Reasoning under uncertainty in deluded schizophrenic patients: a longitudinal study. Eur. Psychiatry 14 (2), 76–83. http://dx.doi.org/10.1016/s0924-9338(99)80721-4.

Cohen, A., Doveh, E., 2005. Significance tests for differences between dependent intraclass correlation coefficients (ICCS). In: Dansereau, F., Yammarino, F.J. (Eds.), Multi-Level Issues in Strategy and Methods. Emerald Group Publishing Limited, Netherlands, pp. 373–420.

Collie, A., Maruff, P., Darby, D.G., McStephen, M., 2003. The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test–retest intervals. J. Int. Neuropsychol. Soc. 9 (3), 419–428. http://dx.doi.org/10.1017/s1355617703930074.

Crump, M.J., McDonnell, J.V., Gureckis, T.M., 2013. Evaluating amazon's mechanical turk as a tool for experimental behavioral research. PLoS One 8 (3), e57410. http://dx.doi.org/10.1371/journal.pone.0057410.

Cumming, G., 2012. Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. Routledge, New York.

Donner, A., Wells, G., 1986. A comparison of confidence interval methods for the intraclass correlation coefficient. Biometrics 42 (2), 401–412. http://dx.doi.org/10.2307/2531060.

Dudley, R., Shaftoe, D., Cavanagh, K., Spencer, H., Ormrod, J., Turkington, D., et al., 2011. 'Jumping to conclusions' in first-episode psychosis. Early Interv. Psychiatry 5 (1), 50–56. http://dx.doi.org/10.1111/j.1751-7893.2010.00258.x.

Dudley, R., Taylor, P., Wickham, S., Hutton, P., 2016. Psychosis, delusions and the "jumping to conclusions" reasoning bias: a systematic review and meta-analysis. Schizophr. Bull. 42 (3), 652–665. http://dx.doi.org/10.1093/schbul/sbv150.

Fisher, R.J., 1993. Social desirability bias and the validity of indirect questioning. J. Consum. Res. 20 (2), 303–315. http://dx.doi.org/10.1086/209351.

Freeman, D., Startup, H., Dunn, G., Cernis, E., Wingham, G., Pugh, K., et al., 2014. Understanding jumping to conclusions in patients with persecutory delusions: working memory and intolerance of uncertainty. Psychol. Med. 44 (14), 3017–3024. http://dx.doi.org/10.1017/S0033291714000592.

Ganster, D.C., Hennessey, H.W., Luthans, F., 1983. Social desirability response effects: three alternative models. Acad. Manag. J. 26 (2), 321–331. http://dx.doi.org/10.2307/255979.

Glass, G.V., 1965. A ranking variable analogue of biserial correlation: implications for short-cut item analysis. J. Educ. Meas. 2 (1), 91–95. http://dx.doi.org/10.1111/j.1745-3984.1965.tb00396.x.

Guilbault, R.L., Bryant, F.B., Brockway, J.H., Posavac, E.J., 2004. A meta-analysis of research on hindsight bias. Basic Appl. Soc. Psychol. 26 (2-3), 103–117. doi:10.1207/s15324834basp2602&3_1.

Hanges, P.J., Lyon, J.S., 2005. Interpreting changes in ICCS: to agree or not to agree, that is the question. In: Dansereau, F., Yammarino, F.J. (Eds.), Multi-Level Issues in Strategy and Methods. Emerald Group Publishing Limited, Netherlands, pp. 421–431.

Hauser, D.J., Schwarz, N., 2015. It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. Sage Open 5 (2). http://dx.doi.org/10.1177/2158244015584617.

Huq, S., Garety, P., Hemsley, D., 1988. Probabilistic judgements in deluded and non-deluded subjects. Q. J. Exp. Psychol. 40 (4), 801–812. http://dx.doi.org/10.1080/14640748808402300.

Jo, M.-S., Nelson, J., Kiecker, P., 1997. A model for controlling social desirability bias by direct and indirect questioning. Market Lett. 8 (4), 429–437. http://dx.doi.org/10.1023/A:1007951313872.

Kerby, D.S., 2014. The simple difference formula: an approach to teaching nonparametric correlation 1. Compr. Psychol. 3. 11. IT. 13.11. doi:10.2466/11.it.3.1.

Kim, H.-Y., 2013. Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. Restor. Dent. Endod. 38 (1), 52–54. http://dx.doi.org/10.5395/rde.2013.38.1.52.

Krug, A., Cabanis, M., Pyka, M., Pauly, K., Kellermann, T., Walter, H., et al., 2014. Attenuated prefrontal activation during decision-making under uncertainty in schizophrenia: a multi-center FMRI study. Schizophr. Res. 152 (1), 176–183. http://dx.doi.org/10.1016/j.schres.2013.11.007.

Lincoln, T.M., Ziegler, M., Mehl, S., Rief, W., 2010. The jumping to conclusions bias in delusions: specificity and changeability. J. Abnorm. Psychol. 119 (1), 40. http://dx.doi.org/10.1037/a0018118.

McGraw, K.O., Wong, S., 1992. A common language effect size statistic. Psychol. Bull. 111 (2), 361. http://dx.doi.org/10.1037//0033-2909.111.2.361.

McLean, B.F., Mattiske, J.K., Balzan, R.P., 2017. Association of the jumping to conclusions and evidence integration biases with delusions in psychosis: a detailed meta-analysis. Schizophr. Bull. 43 (2), 344–354. http://dx.doi.org/10.1093/schbul/sbw056.

Moritz, S., Göritz, A.S., Balzan, R.P., Gawęda, Ł., Kulagin, S.C., Andreou, C., 2017. A new paradigm to measure probabilistic reasoning and a possible answer to the question why psychosis-prone individuals jump to conclusions. J. Abnorm. Psychol. 126 (4), 406. http://dx.doi.org/10.1037/abn0000262.

Moritz, S., Mayer-Stassfurth, H., Endlich, L., Andreou, C., Ramdani, N., Petermann, F., et al., 2015. The benefits of doubt: cognitive bias correction reduces hasty decision-making in schizophrenia. Cogn. Ther. Res. 39 (5), 627–635. http://dx.doi.org/10.1007/s10608-015-9690-8.

Moritz, S., Veckenstedt, R., Bohn, F., Hottenrott, B., Scheu, F., Randjbar, S., et al., 2013. Complementary group metacognitive training (mct) reduces delusional ideation in schizophrenia. Schizophr. Res. 151 (1-3), 61–69. http://dx.doi.org/10.1016/j.schres.2013.10.007.

Moritz, S., Woodward, T.S., Lambert, M., 2007. Under what circumstances do patients with schizophrenia jump to conclusions? A liberal acceptance account. Br. J. Clin. Psychol. 46 (2), 127–137. http://dx.doi.org/10.1348/014466506x129862.

Ochoa, S., Haro, J.M., Huerta-Ramos, E., Cuevas-Esteban, J., Stephan-Otto, C., Usall, J., et al., 2014. Relation between jumping to conclusions and cognitive functioning in people with schizophrenia in contrast with healthy participants. Schizophr. Res. 159 (1), 211–217. http://dx.doi.org/10.1016/j.schres.2014.07.026.

Oppenheimer, D.M., Meyvis, T., Davidenko, N., 2009. Instructional manipulation checks: detecting satisficing to increase statistical power. J. Exp. Soc. Psychol. 45 (4), 867–872. http://dx.doi.org/10.1016/j.jesp.2009.03.009.

Peters, E., Joseph, S., Day, S., Garety, P., 2004. Measuring delusional ideation: the 21-item Peters et al. Delusions Inventory (PDI). Schizophr. Bull. 30 (4), 1005–1022. http://dx.doi.org/10.1093/oxfordjournals.schbul.a007116.

Podsakoff, P.M., MacKenzie, S.B., Lee, J.Y., Podsakoff, N.P., 2003. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J. Appl. Psychol. 88 (5), 879–903. http://dx.doi.org/10.1037/0021-9010.88.5.879.

Portney, L.G., Watkins, M., 2015. Foundations of Clinical Research: Applications to Practice, 3rd ed. FA Davis Company, Philadelphia.

Pos, K., Meijer, C.J., Verkerk, O., Ackema, O., Krabbendam, L., de Haan, L., 2017. Metacognitive training in patients recovering from a first psychosis: an experience sampling study testing treatment effects. Eur. Arch. Psychiatry Clin. Neurosci. 1–8. http://dx.doi.org/10.1007/s00406-017-0833-7.

Rausch, F., Eisenacher, S., Elkin, H., Englisch, S., Kayser, S., Striepens, N., et al., 2016. Evaluation of the 'jumping to conclusions' bias in different subgroups of the at-risk mental state: from cognitive basic symptoms to UHR criteria. Psychol. Med. 46 (10), 2071–2081. http://dx.doi.org/10.1017/s0033291716000465.

Rausch, F., Mier, D., Eifler, S., Esslinger, C., Schilling, C., Schirmbeck, F., et al., 2014. Reduced activation in ventral striatum and ventral tegmental area during probabilistic decision-making in schizophrenia. Schizophr. Res. 156 (2), 143–149. http://dx.doi.org/10.1016/j.schres.2014.04.020.

Rocha, N.B., Queiros, C., 2013. Metacognitive and social cognition training (MSCT) in schizophrenia: a preliminary efficacy study. Schizophr. Res. 150 (1), 64–68. http://dx.doi.org/10.1016/j.schres.2013.07.057.

Ross, K., Freeman, D., Dunn, G., Garety, P., 2011. A randomized experimental

investigation of reasoning training for people with delusions. Schizophr. Bull. 37 (2), 324–333. http://dx.doi.org/10.1093/schbul/sbn165.

Ross, R., McKay, R., Coltheart, M., Langdon, R., 2015. Jumping to conclusions about the beads task? A meta-analysis of delusional ideation and data-gathering. Schizophr. Bull. 41 (5), 1183–1191. http://dx.doi.org/10.1093/schbul/sbu187.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 (2), 420–428. http://dx.doi.org/10.1037//0033-2909.86.2.420.

Tabachnick, B.G., Fidell, L.S., 2007. Using Multivariate Statistics, 5th ed. Pearson/Allyn & Bacon, Boston.

van der Leer, L., Hartig, B., Goldmanis, M., McKay, R., 2017. Why do delusion-prone individuals “jump to conclusions”? An investigation using a nonserial data-gathering paradigm. Clin. Psychol. Sci. 5 (4), 718–725. http://dx.doi.org/10.1177/2167702617698811.

Waller, H., Freeman, D., Jolley, S., Dunn, G., Garety, P., 2011. Targeting reasoning biases in delusions: a pilot study of the maudsley review training programme for individuals with persistent, high conviction delusions. J. Behav. Ther. Exp. Psychiatry 42 (3), 414–421. http://dx.doi.org/10.1016/j.jbtep.2011.03.001.