Acalculia in Parkinson's Disease? A registered report

Hannah D. Loenneker, Inga Liepelt-Scarfone, Klaus Willmes, Hans-Christoph Nuerk, & Christina Artemenko

https://osf.io/qgs5x/?view_only=9ddf45dcfdd846f3998cfb0d842dcf16 version Manuscript_Loenneker_reg_rep_acalculia_PD
Submitted by Hannah Dorothea Loenneker 29 Jun 2021 19:23

Abstract

Elderly people and patients with neurodegenerative diseases such as Parkinson’s Disease (PD) immensely rely on arithmetic skills to lead an independent life. Activities such as medication management, financial transactions or using public transport require intact abilities to manipulate numbers for different arithmetic tasks. However, research on cognitive deficits in PD has been focussing on domain-general functions such as executive functions, attention or working memory so far – largely neglecting potential domain-specific aspects of numerical cognition (e.g. carry or problem size effect). These aspects should be addressed, as PD-immanent deterioration of domain-specific numerical areas and domain-general functions suggests mechanisms of both primary and secondary (mediated by other cognitive deficits) acalculia respectively. The current study will systematically investigate arithmetic performance and effects in PD patients differing in cognitive impairment for the first time, targeting domain-specific cognitive representations of arithmetic as well as the influence of domain-general factors. PD patients with normal cognition (PD-NC) and PD patients with mild cognitive impairment (PD-MCI) will be compared to healthy controls in arithmetic performance in the four basic operations (addition, subtraction, multiplication, division). Discriminant analysis will be employed to assess whether performance in arithmetic tasks can differentiate between a healthy control group and both PD groups. The study results will help us to understand the underlying mechanisms of arithmetic deficits faced by PD patients in daily life.

Keywords: Parkinson’s disease, mild cognitive impairment, arithmetic operation, calculation, place × value system

Round #1

by Zoltan Dienes, 19 Aug 2021 15:04
Manuscript: **https://osf.io/qgs5x/?view_only=9ddf45dcfdd846f3998cfb0d842dcf16**

Dear Zoltan, dear Reviewers,

Thank you for the valuable comments. We revised our manuscript accordingly.

What I really enjoy about Registered Reports is that you can learn from researchers outside your lab and that methodological feedback is implemented before the study will be conducted. The design table suggested by PCI-RR is helpful to clarify thoughts, to differentiate hypotheses and to anticipate unexpected outcomes leading to consequences for theory. So, we are really grateful for this instructive review process so far.

One point we added ourselves is that we had to refine in- and exclusion criteria due to reviewers' comments on the associated registered report addressing basic number processing in Parkinson's Disease (Loenneker et al., 2021). These changes are marked in the manuscript as well.

As the current study is part of a dissertation project that has already been delayed because of the ongoing pandemic, we started recruiting and test the first patient on September 22nd, 2021. Until now, no data has been accessed.

We are looking forward to hearing from you.

Hannah Loenneker
On behalf of all authors

Dear Hannah

I now have two reviewers reports for your manuscript. Like the reviewers I thought your submission was well thought through. The reviewers have some excellent points to make I ask you to respond to in a revision. I also have some points for you to address:

Put the study design table in the manuscript.

We now report the study design in a revised table in the manuscript.

"The difference between PD-NC and PD-MCI cannot be inferred from current literature, which is why we only predict a trend of HC outperforming PD-NC and PD-NC outperforming PD-MCI."

I am not sure what you mean by a trend here. It sounds like you predict two pairwise comparisons. Please clarify. Why not explicitly predict these two, and say for each comparison what theory hangs on the line?

It's true that this statement is imprecise and does not fit our planned analyses. We changed it accordingly:

"Whether and how PD-NC and PD-MCI differ in their arithmetic skills cannot be inferred from the current literature. Therefore, we will further explore the group effect by means of pairwise comparisons between HC and PD-NC and between PD-NC and PD-MCI to identify if arithmetic deficits are more frequent in patients than in controls and in patients with cognitive impairment than in those without."

What you say in the "sampling plan" part of the table is "participants will be tested within a sequential Bayes factor design until the between-subjects factor of group reaches a value of $BF_{10} \geq 6$ or $BF_{01} \geq 6$ for all of the four basic arithmetic operations in research questions (1) and (2)." This doesn't test a trend but an omnibus main effect of group. As an omnibus test it does not precisely test your claims - and requires more subjects than a 1-df test. You could think about two pairwise comparisons as I say.

We agree and changed it accordingly so that the stopping rule does not apply to the main effect of group but to the subsequent pairwise comparisons between HC and PD-NC and between PD-NC and PD-MCI:

"Following the procedure of sample size estimation in Bayes factor design analysis suggested by Schönbrodt and Wagenmakers (2018), participants will be tested within a sequential Bayes factor design until the ~~between-subjects factor of group~~ respective pairwise comparisons between HC and PD-NC and between PD-NC and PD-MCI reach a value of $BF_{10} \geq 6$ or $BF_{01} \geq 6$ for all of the four basic arithmetic operations in ~~research questions one and two~~ Q1 and Q2."

You do not specify a predicted effect size for tests for question 3, nor for covariates in the previous analyses.
You might (or might not) find advice here helpful:
Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science*, 2, 364-377. https://doi.org/10.1177/2515245919876960

One thing I will recommend from that paper is reporting a robustness region for each test, i.e. the set of scale factors that result in the same qualitative conclusion as you reach with your pre-specified scale factor.

Not for all measures, we can infer good effect size estimations and priors, because the study is new and comparable data do not yet exist. There are some related effect size estimations for covariates in Q1 and Q2 and for Q3a based on data from Zamarian et al. (2006), who tested a sample of PD-NC (but not PD-MCI) patients, had a different research question and different target tasks, but at least used the same constructs (but different operationalizations thereof). They reported: "PD patients and controls differed in verbal short-term memory (DOT-serial forward recall, t[41]= -2.18, p <0.04), verbal working memory (digit span backward, t[41]= -3.05, p <0.004), verbal fluency (CERAD-animals and RWT- sports and fruits, p's<0.02), cognitive set-shifting (TMT-B and OMO test, p's<0.04) and interference naming (NAI- FWT, t[41]=2.02, p =0.05). Psychomotor speed was reduced in PD patients (TMT-A and computerized RT task, p's<0.04)."

Whether they found a difference in calculation span tasks depended on the induced working memory load. Patients (*M* = 7.3, *SD* = 1.5) and controls (*M* = 7.9, *SD* = 0.6) did not differ for low WM load, but for medium (patients: *M* = 6.9, *SD* = 2.3; controls: *M* = 7.9, *SD* = 0.2, *p* < .02) and high (patients: *M* = 6.9, *SD* = 2.4; controls: *M* = 7.9, *SD* = 0.2, *p* < .02) loads.

Where they found a difference in complex mental calculation (GDAE) between patients (*M* = 10.2, *SD* = 4.3) and controls (*M* = 15.5, *SD* = 4.6), *t*(41)= -3.70, *p* < .001, they identified associations between the GDAE and interference naming (r = - 0.633, p <0.02), digit span forward (r =0.625, p <0.02) and block span backward (r =0.584, p <0.03). These correlations can be transformed into effect sizes of d = -1.64, d = 1.60, and d = 1.44, respectively. Entering these variables into a stepwise regression analysis on GDAE performance resulted in interference naming being the only significant predictor (β= - 0.63, p <0.02, R^2 = 0.40).

However, many of these tests operationalize the constructs differently (e.g., the GDAE assesses all basic arithmetic operations at once instead of testing the four operations separately as in our study). Even if the name is the same (complex calculation), it is possible that effects depend on the exact stimuli (e.g. with or without carry/borrowing, which we manipulated)..Considerations regarding effect sizes in Q3b cannot be inferred from the literature, as we are not aware of a study comparing PD-NC with PD-MCI patients in mental arithmetic.

With all those limitations in mind, we would like to follow your guidelines on theoretical design from Dienes (2019) and report it in the manuscript as follows:

“We estimate the robustness of the BF analysis based on Zamarian’s results on the Graded Difficulty Arithmetic Test (i.e., mixed arithmetic tasks), compared between PD-NC (M = 10.2, SD = 4.3) and HC (M = 15.5, SD = 4.6), t(41)= -3.70, p < .001 ($M_{\text{difference}}$ = -5.3, SE = 1.43). We assessed the robustness using the online Bayes factor calculator (Dienes, 2018). We defined the likelihood based on a student t distribution with the parameters from Zamarian’s results (M = 5.3, SD = 0.22, df = 41), assuming a Cauchy distribution with the same parameters for the model of the alternative hypothesis and a Cauchy distribution with a location parameter of 0 for the model of the null hypothesis, both one-sided with a lower limit. Results on possible ranges of the scale factors and mean differences producing BFs indicating conclusive evidence are reported in Table 3. These results are only available for ACC, but not for RT data. However, many of these tests operationalize the constructs differently (e.g., the GDAE assesses all basic arithmetic operations at once instead of testing the four operations separately as in our study). Even if the name is the same (complex calculation), it is possible that effects depend on the exact stimuli (e.g. with or without carry/borrowing, which we manipulated). Furthermore, the study was conducted in PD-NC, but not PD-MCI patients. Considerations regarding effect sizes in Q3b cannot be inferred from the literature, as we are not aware of a study comparing PD-NC with PD-MCI patients in mental arithmetic.

Table 3. Robustness considerations of scale factors for Bayesian analysis.

Model for alternative hypothesis			Model for null hypothesis		
Location	Scale	BF_{10}	Location	Scale	BF_{01}
Values based on Zamarian et al. (2006)					
5.3	0.22	376.69	0	0.22	0.003
Manipulation of scale factors with constant location parameter					
5.3	0.022	37808.05	0	0.022	0.00003
5.3	0.7	35.90	0	0.7	0.028
5.3	1	16.81	0	1	0.059
5.3	2.2	3.06	0	2.2	0.33
Manipulation of location parameter with constant scale factor					
0.53	0.22	3.06	0	0.22	0.33
1	0.22	11.96	0	0.22	0.08
53	0.22	38116.94	0	0.22	0.00003

Table 3 indicates that we can expect fairly robust results as long as the location parameters are not substantially smaller, or the scale parameters are not substantially larger than in the work by Zamarian et al. (2006). Since the current study implies both comparisons with a PD-NC and a more advanced PD-MCI group as opposed to a single PD-NC sample in Zamarian's study, we might even expect larger effects. However, we want to be careful with this prediction as target tasks, control variables, and items within tasks differ and may modulate effects. After data acquisition, robustness of the BF across different scale factors will be assessed with a robustness plot in JASP.“

There is still a fair amount of wriggle room for your conclusions for question 3 - please tighten up.

We tried to leave less room for imprecision and interpretational degrees of freedom: (1) by excluding the cognitive covariates from our regression to exclude domain-general effects from our discriminant analysis, and (2) by tightening our possible conclusions and reframing them using Bayesian terminology such as evidence for a certain effect or evidence of absence:

“**Discrimination between cognitive statuses of PD patients by arithmetic performance (Q3).** The last question targeting the diagnostic use of arithmetic will be answered using two Bayesian logistic regressions. This discriminant analysis will be conducted with z-standardized performance in addition, subtraction, multiplication, and division as multiple predictors and with the dependent variable of cognitive status for a) PD-NC and PD-MCI and b) HC and PD-NC. Covariates from ~~Q2~~Q1 will also be included in the model. Both influential case diagnostics and outlier analysis will be applied to minimize the effect of highly influential participants on the regression. The probability for each person to fall into the respective group will be calculated based on the regression. ~~In an additional exploratory analysis, the same procedures will be conducted for the respective arithmetic effects.~~ We hypothesize that the respective group can be predicted with a linear combination of the arithmetic predictors. H3a) The combined performance of the four arithmetic tasks predicts whether a patient belongs to the group with or without mild cognitive impairments. ~~) If the regression model has low values in model diagnostics, arithmetic performance is not suited to discriminate between PD-NC and PD-MCI and arithmetic impairments seem to result from a distinct pathomechanism as opposed to global cognitive impairment. If the regression model has high values in model diagnostics, arithmetic performance is suited to discriminate between PD-NC and PD-MCI and global cognitive and arithmetic impairments seem to share a common pathomechanism H3b)~~ Based on the available literature, we cannot hypothesize whether the HC group will outperform the PD-NC group, despite them both being defined according to a normal cognitive status ~~showing comparable results regarding global cognition.~~ There are several possible unexpected outcomes: (H3a1) If there is evidence for no effect of arithmetic, it is not suited to discriminate between PD-NC and PD-MCI and arithmetic impairments seem to result from a different pathomechanism than cognitive impairment. (H3a2) If there is evidence for an effect of arithmetic, it is suited to discriminate between PD-NC and PD-MCI and cognitive and arithmetic impairments seem to at least partly share a common pathomechanism. (H3b) If arithmetic performance only differentiates HC from PD-NC, but not PD_NC from PD-MCI, it could be used as an early marker for the detection of PD.~~If the HC group outperforms the PD-NC group, arithmetic performance might be used as an early marker for the detection of PD. If both groups cannot be discriminated from each other but PD-NC and PD-MCI do, then arithmetic deficits only occur at a later disease stage.~~

~~.~~”

Don't register your exploratory analyses in Stage 1, i.e. those for which you haven't tied down your analytic options. Just put them in a non-registered section of your results when you submit Stage 2.

Fine, we deleted the parts on exploratory analyses.

Reviews

Reviewed by Pia Rotshtein, 09 Aug 2021 11:11

Loenneker and colleagues proposed to register a study that examine simple arithmetic abilities in two group of patients suffering from Parkinson, those who show no cognitive deficits (PD-NC) and those who show mild cognitive impairment (PD-MCI) in comparison to healthy controls (HC). The study is paired with an additional registered study that examine non-symbolic quantity abilities in the same sample.
Participants will be assess on addition, subtraction, multiplication and division. Accuracy, reaction time and some derivatives measures (differences between trials: borrow and carry over effects) will be the dependent measures. The authors will also consider confounding factors related to socio-demographics and clinical symptoms.

The authors aim to answer three questions: Is PD relates to calculus deficits?; Can tobserved calculus deficits be accounted by core cognitive abilties (e.g. attention, executive function, working memory)? Can calculation problem can be used to detect early PD (PD-NC vs. HC), or emerge at later PD stages (PD-NC vs PD-MCI).

Overall, this is a very large and impressive project. It is very comprehensive and it is clear that the authors have considered the study design very carefully and thoughtfully. The project is exploratory by nature and given it extent and number of measurements it is quite complex.

Thank you very much for the positive evaluation.

Some points to consider
- The main point I want to raise regards the nature of the healthy control group. This is a general comment that probably relates to most studies that uses neuropsychological measures to characterise a specific patient group. the quetsion is who is an appropriate control group that differ only on the factor of interest, here PD.

o The authors should be commented that they aim to recruit carer of the PD patients, which is likely to account for some of the socio-demographic confounds.

Yes, we will recruit partners and caregivers of PhD patients, because we are aware of potential confounds with socio-economic status or other control variables.

“Patient recruitment will be managed through the PD outpatient clinic in collaboration with rehabilitation facilities specialised in PD. Furthermore, PD patients who have been previously studied and gave consent (Ethical vote: 199/2011BO1) to be contacted for potential future study participation will be contacted. The caregivers of the PD patients will also be recruited as healthy controls, in accordance with defined inclusion and exclusion criteria. Additionally, pensioners’ initiatives will be contacted for control group recruitment, with the study being advertised via the university mailing systems.”

o I suggest the recruitment of the control group be more purposeful with the aim of matching them as much as possible with all aspect of the two PD groups.
§ Matching the three groups on age, gender, education, SES and at least childhood testified confidence in math.

We agree that matching our three groups is a crucial topic. For instance, we found gender differences in simple arithmetic before, which might be driven by educational effects in the elderly generation. Therefore, we will match the study groups as exactly as possible (based on the factors you suggest here). However, we need to draw your attention to the challenges we are currently facing in participant recruitment during a pandemic, so we cannot anticipate which level of matching will actually be feasible to include patient groups within the anticipated recruitment phase. We want to address this issue with the following approaches:

(1) We already state in the registered report that groups will be matched by age (max. 5 years mean age difference) and gender (max. 65% male, see section Participants).

(2) We will transparently describe our three groups in a table on sociodemographic and clinical variables, indicating those variables where our groups differed (see section Group-wise characteristics).
(3) We will control for covariates differing between groups at the end of the recruitment process. In the first version of our manuscript, we identified age, education years, gender, Hoehn & Yahr staging, disease duration, and depression as confounding candidates. Based on your suggestions, we will add level of income and educational and professional math experience as further candidates (see sections Group-wise characteristics and Hypothesis testing).
(4) Additionally, if there are differences in other variables like SES, we will use them as covariates in exploratory analyses, to make sure that these are not driving differences between PD groups.
(5) Finally, we address this issue as limitations in the discussion section as regards possible confounds that were not measured.

§ Including two HC groups, one with no mild cognitive impairment and one with MCI is important for arguing that it is the PD that lead to acalculia rather than MCI.

Generally, a research design including a HC-MCI group is a very good idea, but it also leads to some new problems. For instance, it is unclear if MCI and PD-MCI share the same pathomechanism, It has been shown that neurodegenerative cognitive impairments can be caused by brain atrophy, imbalance of cholinergic and dopaminergic neurotransmitters, amyloid pathology (typical of Alzheimer's Disease) and Lewy Body pathology (typical of Parkinson's Disease). Accordingly, MCI is associated to an amyloid and cholinergic pathology as in Alzheimer's Disease, whereas PD-MCI patients display both amyloid, cholinergic, dopaminergic and Lewy Body pathology (Chandra et al., 2019; Lin & Wu, 2015). Therefore it is quite unclear which conclusions to draw if there are interactions of PD diagnosis and MCI status, because in a statistical sense the MCI status is not manipulated in the same way for PD and HC.

While we think, this is an important question for future research, i.e., if MCIs with different underlying pathomechanisms really lead to the same deficits in calculation, it is unfortunately beyond our current resources, because the project is the dissertation project of the first author, who received a grant for it, but we did not get additional project funding. If these studies are successful, we aim at getting a grant, where one important question is, whether different pathomechanisms really lead to the same arithmetic deficits, if one takes a closer and much more detailed look at mathematical performance as we wish to do.

For these reasons, we focus on the two comparisons, which deemed most important for us as a starting point: The group effect in our main analyses will be further differentiated in pairwise comparisons between HC and PD-NC on the one hand and between PD-NC and PD-MCI on the other hand. Consequently, the difference between HC and PD-NC is the diagnosis of Parkinson's Disease, with both having normal cognition. To identify if cognition irrespective of the PD related disease progression affects arithmetic performance of PD-NC and PD-MCI will be compared (see for example research questions 2 and 3a). We specified these considerations on research design in our manuscript (see section Hypothesis testing).

We also added this line of argumentation to the limitations section:

“Our study has an incomplete design with only one matched healthy control group. Thus, arithmetic deficits in the PD-MCI group are difficult to interpret as both general effects of cognitive impairments and/or specific disease progression potentially contribute to this effect. To clarify this issue, future studies can include an additional control group with non-PD related MCI (neglected here due to limited resources). Moreover, it is unclear if MCI and PD-MCI share the same pathomechanism, and therefore lead to the same deficit. Neurodegenerative cognitive impairments can be caused by brain atrophy, imbalance of cholinergic and dopaminergic neurotransmitters, amyloid pathology (typical of AD) and Lewy Body pathology (typical of PD). Accordingly, MCI is associated to an amyloid and cholinergic pathology as in AD, whereas PD-MCI patients display both amyloid, cholinergic, dopaminergic and Lewy Body pathology (Chandra et al., 2019; Lin & Wu, 2015). Therefore, interaction of PD patient status and MCI may be due to particular underlying pathomechanisms, which may be unknown and/or heterogeneous for an HC-MCI group. As a start, we will compare HC and PD-NC on the one hand to identify effects specific to PD (both with normal cognition), and PD-NC and PD-MCI on the other hand to investigate the effect of cognitive impairment within PD. Based on these results, it seems worthwhile to investigate whether MCIs in different groups with different pathomechanisms lead to the same arithmetic deficits or not in future studies.”

- If I understand correctly the upper limit for participants number is 120.

o Will these be equally divided across all three groups?

Yes, that’s true. We made this fact more explicit (see section Statistical power analysis and sample size estimation):

“Due to feasibility, an additional maximum sample size of n_{max} = 120 valid data sets is established (targeting equal group sizes).”

o Could you provide a power analysis for 120 participants for a medium effect size, as this is the maximum power you can be achieved.

A frequentist statistical approach would require power analysis. However, for the Bayesian approach, we instead implemented a sequential Bayes factor design for sample size calculation. This is why we opted for the Monte Carlo simulations for a group difference with a medium effect of d = .5:

“We estimated the properties of the planned research design with Monte Carlo simulations as implemented by Schönbrodt and Wagenmakers (2018) based on a sequential boundary of BF_{10} = 6, d = 0.5, n_{min} = 35, n_{max} = 120 and 10,000 simulated studies. Simulating the performance of our design under H_1 resulted in 9% of studies terminating at n_{max}, 90.8% terminating at H_1 boundary and 0.2% at H_0 boundary, on average stopping at n = 63. Simulating the performance of our design under H_0 resulted in 35.6% of studies terminating at n_{max}, 3.6% terminating at H_1 boundary and 60.7% at H_0 boundary, on average stopping at n = 103.”

This means that we chose a maximum sample of *N* = 120 (i.e., 40 per group) because of feasibility considerations. Testing can stop earlier, if all effects of interest reach the conclusive criteria of BF_{10} = 6 or BF_{10} = 1/6. If testing stops because we reached the maximum sample size and some effects of interest have BF_{10} values between 1/6 and 6, these results remain inconclusive and need to be addressed in follow-up studies.

o It can be that when considering all the additional covariates 120 is too small group.

As described above, we had to set a maximum sample size of 120 because of feasibility reasons. According to Lakens (Lakens, 2021) limited resources are an appropriate consideration when estimating sample size. Note that this sample size might still be too small even though it is between 2 and 3 times larger than the samples in the studies we used to estimate the empirical effect size. However, the Bayes factor still allows to interpret how likely the hypothesis is given the data. As we're conducting the first study investigating arithmetic deficits in Parkinson's Disease in this way, our evidence can be used to inform future studies focusing on a specific effect and conducted with an even larger sample. We added this issue to the discussion of possible limitations.

- The analysis suggests to exclude participants who got the first 10 question wrong in one of the tasks. This may dilute your effect. Maybe in the practice session include some very simple arithmetic to ensure that they understand the task instruction. Then if they understand the task but nevertheless fail the first 10 questions you can just award them a zero score.

o Will the first 10 be the easier examples, less complex?
o Will participants be asked if they want to stop or continue?
o Will participants received feedback on their answers?

Thank you, this is a very good point. To clarify our procedure: each arithmetic test begins with 10 practice items, which can be repeated to ensure understanding of task instructions. After this, the first ten experimental trials are counted and the experiment stops if participants cannot solve any of these. Trials are presented in randomized order, and not with gradually increasing difficulty.

As suggested, we will assign a value of 0 for ACC if the experiment was stopped because of 10 wrong answers at the beginning of the experiment. These participants will be excluded from the RT analysis (since it is based on correctly solved trials only) but still be used for an additional ACC analysis to see whether the results change, if such patients are included (because it may be unclear, if they cannot do specific processes targeted by the task (e.g., because of domain-specific numerical deficits) or because they are not able to memorize the instruction (because of more domain- or task-overlapping) verbal or memory. Practice trials will consist of easier arithmetic problems (e.g., using the operand 1 and single-digit numbers).

Participants will not receive feedback in order not to frustrate them. For the same reason, they will not be told that the experiment was stopped because of 10 wrong answers but it will look as if the experiment stopped normally.

- The number of planed tests is huge. It is good you opted for Bayes factor as you will not need to compare for multiple comparison. However, if only a single of these test will results in significant results – what would you conclude on the relation of PD and calculus processing?

To make this clear – our stopping rule applies only if all effects of interest reach the conclusive criterion of $BF_{10} > 6$ or $BF_{10} < 1/6$. If – unexpectedly – only a low number of effects provide conclusive evidence for or against a group difference, we will only interpret these specific components and recommend studies with larger sample sizes for future research to depict this seemingly complex mechanism of degeneration. In the current study, we only aim to provide an overview on basic arithmetic operations in PD for the first time. The findings might serve as a starting point for future research focusing on specific factors (e.g., verbally mediated fact retrieval) or using a large-scale design (e.g., multi-lab setting). We added this limitation to the manuscript.

- When computing the 'borrow effect' could you please provide a clear definition of problem size.

Thanks for requesting this clarification. We reworked this part as follows:

“This allows for calculating the borrow effect, defined as the difference in mean RTs between borrow and non-borrow problems, counterbalanced for problem size (which is defined as the value of the first operand, i.e., the minuend).”

- Could you please provide a clear definition of the task complexity factor that will be used to answer Q2.

Sorry for the inconsistent use of terminology. We introduce the definition of item difficulty in the introduction:
“Where carry- and borrow-effects define task difficulty in addition and subtraction, difficulty in multiplication and division is defined respectively based on problem size of the product or divisor, with the problem size effect showing faster RT and higher ACC for smaller (i.e., easier) problems (Domahs et al., 2006; Zbrodoff & Logan, 2005).”

As the factor difficulty was not described clear enough regarding our research design, we added this definition in the methods section for each basic arithmetic operation (addition: carry, subtraction: borrow, multiplication and division: problem size).

- Given the number of potential covariates it maybe more useful to remove variability of socio-demographic and clinical factors from the tasks performances before moving into Q2 (impact of core functions). Then removing the additional variability of cognitive covariate from task performances before using the data to answer Q3 (ability of calculus to discriminate between the three groups).

We hope that we understood your question correctly: Do you suggest to conduct a linear regression for Q1 with the potential covariates and then predict new covariate-free values for the dependent variable using the regression equation? And then continue on the covariate-free values for Q2: first conduct a new regression with potential cognitive covariates, predict cognitive-covariate-free values and use them to predict the dependent variable in Q2 and further also for Q3? We are not sure whether your rationale for this approach was to reduce the number of covariates in analyses for Q2 and Q3 to counteract running into a power issue, but this procedure probably would not solve the problem.

Furthermore, this procedure would not fit our planned analysis with Bayesian ANCOVAs, being a different class of analysis. In our analyses, we do correct for the influence of these factors as covariates in each question, which actually should show the same results as above if we correctly understood it. We also added the following sentence to the limitations section: “Future standardized assessment of arithmetic skills should establish norms correcting for confounding factors we identify in the current study such as age, education, or gender.”

We are aware of the important role of SES and would consider them in the exploratory analysis. For instance, one could regress our data against SES and conduct the proposed analysis for the residuals to explore if there are any substantial changes, when SES is partialed out. As we were told not to describe exploratory analyses in the manuscript, this point will only appear in the Stage 2 Registered Report in case our study gets accepted for publication.

- Could you please provide a measure that will ensure that the quality of the data is high (e.g. expecting to replicate basic accuracy and RT effect on task complexity, independent of PD, or across all groups)

That’s a very good idea. We added the following manipulation check: Before hypothesis testing, we will analyze the carry effect for addition, the borrow effect for subtraction, and the problem size effect for multiplication and division in the HC group regarding ACC and RT. Replicating these typical arithmetic effects will enable us to check the data quality in our task. This will be done with Bayesian *t*-tests.

- dividing the text by sub-headings, will aid the reading and comrehension of all section. using some additional tables to summarsie the information (description of tasks, measures collected) can also help.

We reworked the text, by shortening it and introducing subheadings. We further added a table summarizing the measures conducted (Table 3). We hope this increases readability.

Reviewed by anonymous reviewer, 19 Aug 2021 10:02

Thank you for asking me to review this detailed, well written and interesting RR. The research question has good validity and the rational of study as well as hypothesis make sense. The sample size and inclusion/exclusion criteria are well defined and justified. Overall, the analysis is feasible and appropriate to answer the questions. However, I have identified the following concerns that warrant expanding/clarification:

- I understand the RR will be ran alongside the other similar study already accepted-in-principle. Can you give us details on how task order and administration of both studies will be managed and lay out clearly what data overlaps between the two studies?

We described the tests conducted for the associated registered report in the procedure section. These are limited to the basic numerical tasks transcoding, number line estimation, non-symbolic magnitude comparison, symbolic magnitude comparison. The question of this first study was, whether PD patient and control groups differ already regarding basic foundations of number processing, such as the magnitude representation, their verbal representation or their spatial representation (Loenneker et al., 2021).

In the current study, we systematically focus on arithmetic, which needs to be done by humans almost on a daily basis (e.g., financial matters or even fundamental things such as cooking with recipes or checking the timing). These arithmetic tasks (addition, subtraction, multiplication, division) are exclusively investigated and reported in the current report. They are less elementary, more complex and eventually more linked to daily activities (where you rarely are asked to estimate numbers on a spatial line, which is an important fundamental task for space-number representations). None of the arithmetic tasks is reported in the associated registered report. All other measures (clinical, demographic, cognitive variables) will be collected and analysed in both studies:

“As part of a broader research project on numerical cognition in PD, the current study will be conducted in joint sessions with the registered report that investigates basic number processing in PD, which has already been granted in-principle acceptance (for further measures conducted regarding basic number processing in PD see Loenneker, Artemenko, et al., 2021). After obtaining written informed consent, the predefined inclusion and exclusion criteria will be used to assess participant eligibility in a semi-standardised questionnaire. PD patients’ cognitive performance will be used to assign them to the PD-NC or the PD-MCI group. Participants will attend two sessions of 1.5 to 2 hours each. In order to handle patient attrition, there will be breaks within each session as required. The first experimental session consists of the sociodemographic questionnaire, the basic numerical tasks (not considered here: transcoding, number line estimation, non-symbolic magnitude comparison, symbolic magnitude comparison, Loenneker, Artemenko, et al., 2021) and the numerical arithmetic tasks addition, subtraction, multiplication and division in this order. Afterwards, participants and caregivers will complete the clinical scales and questionnaires which may also be filled out at home between the first and second session. In the second session, the MoCA, clinical variables, motor assessment, and neuropsychological test battery will be conducted in that order. The two sessions may be scheduled three weeks apart at a maximum. On a conceptual level, the first publication (Loenneker et al., 2021) addresses the basic foundations of number processing, whereas the current publication focuses on arithmetic skills, which are more relevant for a patient’s daily life.”

- The introduction is quite long and needs to be more succinct focusing on the question at hand and leading to research question and hypothesis. Its missing a good overview of PD and its epidemiology and symptoms. Its not common to have hypothesis in supplemental so I would incorporate this within main text.

We restructured and shortened the introduction for increased readability and to guide the reader to the research question in a straightforward manner. We further reworked the section on PD and added a respective sub-heading to introduce epidemiology and specific symptoms.

As the table was required from PCI-RR, we initially added it as a supplementary file. We now added it to the main document, but in a shortened and reformatted way (Box 1).

- Within the arithmetic tasks however you propose to analyse RT. However, the response is entered by experimenter. Could you use a microphone instead and record verbal response? Otherwise, your RT measure will not be very reliable.

We are sorry that we described our experimental paradigm not well enough that it can be precisely understood. We clarified it accordingly:

“Every trial starts with a fixation point in the shape of “o” for 750 ms and is followed by the arithmetic problem presented centrally on the screen until the participant presses the space key while responding orally. The participant starts pressing the space key when starting to answer, holds it in the meantime, and releases the key when having finished answering. The critical RT here is the first key press, and the time between key press and key release will be used to exclude participants taking too long due to utterances intermixed with the answer. The numerical response is entered with a QWERTZ keyboard by the experimenter, who then initiates the next trial (for a similar procedure in elderly see Artemenko, 2021). By decreasing motor effort for PD patients, this response format aims at minimizing the influence of PD-immanent motor impairments on arithmetic performance while logging both response and RT.”

This procedure will be trained in the practice trials where easier tasks are presented and where the instructor can insist on proper execution of the task.

- The authors should a section on limitations of the study and how these will be dealt with regarding design, analysis, and previous findings. Issues covered could include for example issues such as heterogenous clinical sample, fatigue, diagnosis errors. Finally a section on potential unexpected outcomes should also be included including analysis or other steps taken to overcome these.

As there were also some concerns by the other reviewer, we added a section on possible limitations and unexpected outcomes:

“Several issues might be encountered in the following stages of recruitment, testing, data analysis and interpretation.

Due to the potential problems in patient recruitment due to pandemic restrictions, group matching might not be as successful as intended. Cognitive impairment in PD is of a heterogeneous nature. Patient groups might, for example, differ regarding disease duration, PD motor type, or non-motor burden. By controlling for the main confounding variables, we at least partially account for the heterogeneity in our group comparisons. Future standardized assessment of arithmetic skills should establish norms correcting for confounding factors we identify in the current study such as age, education, or gender. Due to possible confounds, the number of covariates might exceed the statistical power needed to detect effects within the given sample size. Therefore, the maximum sample size of 120 might still lead to an underpowered study, even though it is 2-3 times larger than the samples in the studies used for effect size estimation. However, the Bayes factor still allows to interpret how likely the hypothesis is given the data. As we are conducting the first systematic study investigating arithmetic deficits in PD, our evidence can be used for sample size calculations in future studies focusing on a specific effect and conducted with an even larger sample.

Cognitive diagnosis of PD-NC and PD-MCI will only be made based on MDS Task Force level I criteria – and not based on a comprehensive level II neuropsychological test battery, because testing time is limited and we want to reserve enough time for the target tasks of interest. This compromise might increase the probability of a misdiagnosis of a patient's cognitive status, because reliability and validity of level I criteria might be lower. To inform about cognitive test performance, the number of patients scoring ≤ 1.5 standard values below the population mean as specified in the respective test manual will be reported.

To reduce patient attrition, the testing is split into two experimental sessions, including enough breaks for the patients to recover and they will be allowed to take their medication during the session. We will conduct the tests in a standardized order, which might induce group differences if the three groups are differently affected by attrition. However, we can reduce variance in the effects that the different tests have on one another. Although attritional and motivational effects are reduced by splitting the testing up into two sessions, it is still possible to have some remaining effects within one session.

Our study has an incomplete design with only one matched healthy control group. Thus, arithmetic deficits in the PD-MCI group are difficult to interpret as both general effects of cognitive impairments and/or specific disease progression potentially contribute to this effect. To clarify this issue, future studies can include an additional control group with non-PD related MCI (neglected here due to limited resources). Moreover, it is unclear if MCI and PD-MCI share the same pathomechanism, and therefore lead to the same deficit. Neurodegenerative cognitive impairments can be caused by brain atrophy, imbalance of cholinergic and dopaminergic neurotransmitters, amyloid pathology (typical of AD) and Lewy Body pathology (typical of PD). Accordingly, MCI is associated to an amyloid and cholinergic pathology as in AD, whereas PD-MCI patients display both amyloid, cholinergic, dopaminergic and Lewy Body pathology (Chandra et al., 2019; C. H. Lin & Wu, 2015). Therefore, interaction of PD patient status and MCI may be due to the particular underlying pathomechanism, which may be unknown and/ or heterogeneous for an HC-MCI group. As a start, we will compare HC and PD-NC on the one hand to identify effects specific to PD (both with normal cognition), and PD-NC and PD-MCI on the other hand to investigate the effect of cognitive impairment within PD. Based on these results, it seems worthwhile to investigate whether MCIs in different groups with different pathomechanisms lead to the same arithmetic deficits or not in future studies.

Finally, it is possible that we do not find any group effects, but all patients show the same performance as the elderly control group. Firstly, this could mean that impairments of PD patients observed in clinical practice have to be attributed to mere aging effects or that arithmetic in everyday life is even more complex than our experimental manipulation. In this case, we would recommend future studies that either compare performance to a young control group or to a PDD sample which previously has been shown to have arithmetic deficits (Kalbe, 1999), or employing more complex arithmetic tasks. However, it could also be a result of aggregating data across individuals. Another approach from differential psychology might solve this issue by looking at the different arithmetic effects on an individual level (i.e., proportion of individuals who show a certain effect or deficit per group) or using LMMs, which, however, would require much higher power to be conclusive. Finally, missing group effects could be a consequence of ceiling or floor effects, either at the item or the task level – however, since we are one of the first manipulating arithmetic item difficulty substantially within tasks, we are optimistic that we will not have ceiling or floor effects throughout all conditions."

Minor issues

'the magnetic resonance imaging volume of the angular gyri' – this is unusual language here and instead I would remove MRI and just say cortical volume?

Thanks, we changed the sentence accordingly:

This link between arithmetic skills and daily functioning is further supported by findings that the cortical volume of the angular gyri (involved in arithmetic fact retrieval) predicts financial deficits in MCI (Griffith et al., 2010).

A few typos noted (e.g., page 5 end paragraph, line 7).

We revised the manuscript and deleted all errors detected. We hope that we found all typos in the article.

Spell out acronyms first time within text (e.g., ADL; MOCA)

Thanks for pointing us to this issue. We hope that we properly introduced all acronyms now.

Some of the paragraphs are long and should be split into separate paragraphs (e.g., page 3).

Thanks for this comment. We reworked the manuscript to make it more concise.

Reviewed by Ann Dowker

Recommendation:

I am very happy to endorse this submission. The study is very important. As the authors say, there has been a lot of research on domain-general cognitive deficits in Parkinson's disease, but much less on specific numerical deficits. It is important that these should be investigated; and the study to test this is very well designed and well planned.

It is also important to study the extent to which any differences in numerical cognition between individuals with Parkinson's disease and others are truly domain-specific or secondary to deficits in executive functions. This is now seen as an important issue in many other areas of numerical cognition: e.g. in typical development; in children born preterm; in children and adults with mathematical difficulties; in individuals with developmental disorders such as ADHD and ASD. It is clearly important to study this issue in people with Parkinson's disease; and the authors have designed their study very well for this purpose.

I am very impressed by the authors' literature review, which is clear, comprehensive, and embeds the study very well in the previous research background.

The planned statistical analyses appear highly appropriate and suitable tests of the findings.

My only real criticism is in fact of the title! I think that the term 'acalculia' could be replaced by 'specific numerical deficits' or similar; as it is not certain that the study will find deficits of the severity usually described as 'acalculia', and the findings are likely to be interesting and important even if they do not.

Thanks a lot for these kind and supportive words. We changed the title to "arithmetic deficits" to contrast this study with our associated registered report on basic numerical deficits (Loenneker et al., 2021). We further addressed this change of terminology in the main text of the article as follows:

Domain-general functions such as attention, working memory, or executive functions are an additional prerequisite for successful ~~number~~ arithmetic processing and deficits therein can give rise to arithmetic deficits comparable to secondary acalculia (termed secondary arithmetic deficits hereafter, Knops et al., 2017).