Review 1

Dear Dr. Kräplin,

Thank you for submitting your Stage 1 Registered Report, entitled "Impulsivity and online sports betting behavior: Untangling the causal relationship" to PCI RR. I have now received comments from two expert reviewers, and have also read the manuscript myself. Combining player tracking data and survey data provides a unique opportunity to understand the etiology of gambling disorder, which is also a key strength of the proposed research. However, my major concern with the current manuscript is that certain methodological details do not seem to be fully specified, which can leave room for potential undisclosed flexibility in the procedures (see the review criteria 1C and 1D on the website of PCI RR). The reviewers and I have noted several aspects where more clarity is needed to reduce such potential flexibility. I hope these comments will be useful to you, and I look forward to seeing a revised version of the manuscript.

Dear Dr. Zhang Chen,

We would like to thank you and the reviewers for your careful reviews and helpful feedback. We have improved our paper in terms of clarity and address other concerns. Our paper is now much improved and we are happy about resubmitting it for further review. Please note that the lines referred to in the reply correspond to the manuscript in track change mode (not the original manuscript).

Please provide more information for the power analysis, such as: the exact statistical model used in the simulation, the assumed effect sizes (at the moment, the text only says "small associations" and "a small to medium association", without giving any specific values), the criterion for statistical inference in the simulation (an alpha level of .05?), and the observed power for each of the three hypotheses (I assume they will differ for the three hypotheses?). Such information will help readers better appreciate what effect sizes the target sample size will have a reasonable chance of detecting.

We agree that it is important to provide more detailed information about our power analysis. We now explain this aspect in more detail in the section “Sample size and power calculations”:

‘We calculated the required sample size for the study based on Hypothesis 3, as it was expected that the mediation hypothesis would require the largest sample size. In order to determine the appropriate sample size for our study, we relied on a simulation study that specifically explored the sample size requirements for testing within-person indirect effects in longitudinal data (Pan et al., 2018). In their simulation study, they defined a small effect size as 0.14, ‘halfway’ as 0.26, medium as 0.39, and large as 0.59. We assumed an intra-class correlation of GD during 9 months of 0.6 (Currie et al., 2013), small associations (0.14) between impulsivity and risky gambling behavior during 9 months (based on a study of multiple addictive behaviors; Kräplin et al., 2020), and a small to medium association (0.26) between risky gambling behavior and GD during 9 months (Xuan & Shaffer, 2009). The required sample size for our study was calculated to be n = 370 using the bootstrap method and given the planned three waves, a planned power of 80%, and an alpha level of 5% (see

Table 4; Pan et al., 2018). Since Hypotheses 1 and 2 require a smaller sample size than Hypothesis 3, the sample size determined for Hypothesis 3 provides sufficient statistical power to adequately test Hypotheses 1 and 2.'

Both reviewers have concerns about whether the target sample size of 370 can be reached given the sample size at the first wave. This is of course an inherent difficulty with such a longitudinal design - one can never know in advance what dropout rate will eventually be observed. That the initial sample size at wave 1 slightly exceeded the planned sample size, and the dropout rate observed in a past study with a similar design to some extent support the feasibility of achieving the target sample size at wave 3. It would nevertheless be good to know if there are any contingency plans in place to address this potential issue.

We also consider it very important to achieve the target sample size. Due to your comment and those of the other reviewers, we realized that we did not explain in enough detail how we calculated the sample size. We have revised the section 'Sample size and power calculations' and gave more detail concerning the estimated response rates, the retention rates, and the data exclusion rates as following (lines 183 – 195):

'Based on the experience from the RIGAB 12 month follow-up, a retention rate of 60% is expected between the first online survey and the third survey wave 9 months later. In addition, the exclusion of further data due to data quality checks was estimated at 25 %. Based on these response rates, the retention estimates, and the data exclusion estimates, 925 participants need to participate in the first wave to achieve a sample of 370 participants in the third and final wave (60% retention rate, 25% data quality exclusion, 10% buffer). Assuming a linear decline of the retention rate, we estimated that 555 participants have to participate in the second wave (80 % retention rate and 25% data quality exclusion). Therefore, we estimated that we would have to invite 10,153 Tipico account holders in the first wave, of whom about 1,320 would be assumed to respond (response rate 13%) and n = 925 (30% data quality exclusion) could be included in the final sample of the first wave.'

Furthermore, we would like to address the issue of retention rate in three ways. The first way is also described in the answer to your next point. We have planned very carefully, how to achieve a high retention rate based on previous experience (see the response to your next point for more details). Second, we are right now collecting data for the second wave from which we can conclude that our calculations of the retention rate are reasonable. Specifically, we have already re-contacted 487 participants and 352 have participated in wave 2, which translates to a retention rate of 72.3% (80% was planned). Two more reminder emails will be sent. Due to data quality, 16 % were excluded (25 % were planned). In sum, we are therefore confident that we will achieve the target sample size in wave 3. Third, we will treat the missing values with full information maximum likelihood (FIML). FIML can provide unbiased inferences and valid standard error estimates in the case of nonresponses. We have now included all this information in the manuscript. The first and second points have been included in the 'Recruitment and retention' section (line 220 ff.) and the third point in the 'Statistical analysis' section (line 529).

Related to the previous point, since data collection for wave 1 is already finished, achieving a low dropout rate for waves 2 & 3 seems to be key here for reaching the target sample size.

More methodological detail on the recruitment procedure would be helpful in this regard. For instance, how much time did/would participants have to participate in the task after getting an invitation? Did/would participants get reminders if they did not respond to the initial invitation, to potentially reduce the dropout rate for waves 2 and 3? In general, more details could be provided for the recruitment procedures for all three waves, to give a better picture of how the recruitment took place/will take place.

For us, sample retention rate is also a key factor in longitudinal studies. We have integrated several strategies into the study to minimize the dropout rate:

- Increasing amounts of recompensation for participants over time, with the highest amount of 30 euros paid out in wave 3 for quite a short participation time (15 minutes).
- In contrast to the RIGAB study, we have decided not to use vouchers, as this reduces the likelihood of participation (and compared to the RIGAB study, we are not recruiting particularly vulnerable participants for whom vouchers would be less risky because they cannot directly reinvest the money in gambling).
- We communicate regularly with participants by email, for example if bank details for the participation recompensation are incorrect.
- We also send three reminders every 2 weeks (with the option to opt out), giving participants a total of 8 weeks to complete the study.

We now describe these strategies in more detail in the manuscript under the subsection 'Recruitment and retention' (line 224). In addition, we included the new subsection 'Procedure' to better explain the procedure for all three waves (line 373 ff.).

Please specify how the balanced integration score will be calculated. I think it will be more convenient to the readers, and also reduces potential flexibility in data analysis, if they do not need to look up another source to understand this method.

We agree that it is more convenient to shortly explain the calculation of the balanced integration score. We now have added the following explanation:

'The BIS is calculated by first standardizing the reaction times (correct Go-Trials) and the percentage of correct rejections (correct Nogo-Trials) to bring them to the same scale. Then the standardized reaction times are subtracted from the standardized percentage of correct values."

Please also specify how the scores for impulsive personality traits will be computed. For instance, will the sum score, average score, factor score (or an alternative score) be used for each of the three factors?

We now clarify that the summed values are used to calculate each of the three factors of the impulsive personality traits, i.e. the summed value of the 'positive urgency' scale and the summed value of the 'negative urgency' scale are added to the 'urgency' factor. We have now included this information in the subsection '(c) Impulsive personality traits' (line 310).

More clarity is needed for how the variables operationalizing riskier gambling behavior will be constructed. For instance, for "a higher variance (mean standard deviation) in the number of bets and stakes over 3 months", will the number of bets/stakes first computed per day, or per week, or per month, before the standard deviation is computed? For "betting at higher odds (mean weighted odds) over 3 months", how will the odds be weighted exactly? For "no or low gambling limits set by the player over 3 months", what constitutes "low limits" in this case?

We see that we need to clarify the operationalization of risky gambling behavior. To facilitate the understanding of our operationalization and to increase transparency, we have now added Table 2 on page 14 where we show which data will be received from Tipico and how exactly we will calculate the respective variables for analyses. Please note that we skipped the variable 'gambling limits' as the RIGAB study showed that a very low number of account holders ($n < 10$) is using such limits.

To promote transparency, please consider sharing the survey used in this project. Please also describe the procedure in the survey for all three waves (e.g., in which order do participants receive the different tasks and questionnaires etc.). Will the same plausible time limit (i.e., 10 minutes) also be used for waves 2 and 3?

We are happy to share the survey on PsychArchives under https://doi.org/10.23668/psycharchives.14178. Please note that the study material is only available in German. We now describe the order of the different tasks and questionnaires and take all three waves into account when describing the survey in the new subsection 'Procedure'.

Line 354: "we expect some individuals with too few data points to compute a reliable aggregated variance". What criterion will be used to say that a certain individual has "too few" data points?

We now have added the following explanation (line 446):

'A minimum number of at least one bet / deposit / withdrawal per month (i.e. a total of 3) is required for the calculation of the pooled standard deviations (Table 2). For individuals with less than three bets/ deposits/ withdrawal, there will be missing variance sum values.'

As one reviewer pointed out, please specify how the number of factors will be determined for the variable risky gambling behavior. At the moment, the description (lines 372 - 377) seems to largely rely on subjective judgment, which leaves much room for flexibility in data analysis.

We now describe more transparently how we will select the number of factors to operationalize risky gambling behavior. Factors will be selected using the eigenvalue, scree

plot, parallel analysis, and interpretability. A detailed description of our selection criteria has been added to the subsection ‘Operationalization and study materials’ (lines 351 – 366).

For H1, five separate models will be run. For H2 and H3, potentially more models will be run, depending on the number of factors obtained for the variable risky gambling behavior (which needs to be further specified, see above). This raises the question of how the results of these separate models will be used to infer whether a certain hypothesis is supported or not. For instance, will H1 only be considered to be supported when all 5 models reveal an effect of impulsivity on gambling disorder severity, or will the different facets of impulsivity be considered separately? If it's the latter case, the hypotheses will also need to be adjusted accordingly to reflect that the statistical inference will be made for each facet (instead of impulsivity as one construct). The same applies to H2 and H3.

We thank the reviewer for these very important questions, which have encouraged us to better describe how potentially significant results from our separate models are used to infer whether a particular hypothesis is supported. We added in line 118 of the introduction that *‘we assume in detail that the predictor impulsivity will consist of (a) impulsive choice, (b) impulsive action, and (c) impulsive personality traits.’* In the subsection ‘Hypothesis testing’, we added in detail under which circumstances we will consider a hypothesis to be confirmed. For example for Hypothesis 1:

‘To test Hypothesis 1a, the predictor is impulsive choice (k value), to test Hypothesis 1b, the predictor is impulsive action (BIS), and to test Hypothesis 1c, the predictors are impulsive personality traits (urgency, lack of conscientiousness, and sensation seeking). […] Hypothesis 1c will be confirmed if two of the three personality factors significantly predict GD. Overall, Hypothesis 1 will be fully supported if all sub-hypotheses are confirmed, and partially supported if at least one sub-hypothesis is confirmed.’

Multiple indices will be used to assess model fit. Please specify how these indices will be used in the statistical inference. For instance, will the results from a model only be interpreted if all (or if any) of the indices show that there is adequate fit? Differences in model fit will also be computed, but it is not entirely clear to me which models will be compared, and what conclusions will be drawn from such model comparisons.

We have updated our fit thresholds to provide clear cut-points for adequate and good fit. Please see lines 531 – 540 in the manuscript: *‘Model fit of the RI-CLPM will be evaluated using three indices: the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). For CFI, values larger than .95 indicate good fit, while values between .90 and .95 indicate adequate fit. For RMSEA and SRMR, values below .05 indicate good fit, while values below .08 indicate adequate fit (Gunzler et al., 2021; Kline, 2015; West et al., 2023). Model results will only be interpreted if all fit indices demonstrate good fit, or if the indices show a combination of good and adequate fit.’*

The final sentence summarizes our approach to collectively using these fit indices to determine if a model is interpretable or unacceptable.

The sentence about the differences in model fitting was mistakenly included in the manuscript. We have removed the sentence and apologize for the error.

Line 280: "lack of consciousness" should be "lack of conscientiousness"?

Yes. It has been corrected.

Review 2

I have had the opportunity to review Stage 1 Registered Report 574: "Impulsivity and online sports betting behavior: Untangling the causal relationship". There are potential merits to this study, especially the synthesis of gambling records with survey data, which remains an important unexplored method in gambling studies. However, I have noted several issues with the planned research, including one major issue regarding the study's ability to achieve a large enough sample size by Wave 3 to ensure adequate power for conducting the main analyses. I would entertain a rebuttal, or otherwise any other assurance, from the study's authors regarding this issue. Pending this type of assurance, however, I am hesitant to recommend signing off the current research plan as it stands.

We would like to thank you for your careful review and constructive feedback. We were happy to improve our paper in terms of clarity and address other concerns. Please note that the lines referred to in the reply correspond to the manuscript in track change mode (not the original manuscript).

Major

Reading through the cited study (Czernecka et al., 2023), I assume the authors are referring to this particular result:

"Of the 555 potential participants, 325 participated in the follow-up (58.5%)".

If we we're to expect this type of retention rate based on the baseline n = 954 across both subsequent waves, then we would expect approximately 558 respondents by Wave 2 and approximately 326 respondents by Wave 3, which is below the 370 respondent threshold identified in power analysis. There's a lot to consider with this estimate, however. First off, what happens when we add the 30% reduction to both waves based on data cleaning? Conversely, would we expect the retention rates and data quality to increase at wave 3, as the study is self-selecting more willing and quality respondents at Wave 2? In any case, there remains the distinct possibility the study will be underpowered at Wave 3, and the study authors have not provided any sort of contingency plan for how they plan to address this possibility. This issue in particular makes me hesitant to greenlight this study for Stage 2 until I've seen the result of final (wave 3) data collection, or received some other adequate assurance/rebuttal to this issue.

Thank you for bringing up this issue. We also consider it very important to achieve the target sample size. We would like to point out that we mistakenly referred to the retention rate of 60% in the RIGAB study as the 'dropout rate' in the paper, which we have now corrected. Due to your comment and those of the other reviewers, we realized that we did not explain in enough detail how we calculated the sample size. We have revised the section 'Sample size and power calculations' and gave more detail concerning the estimated response rates, the retention rates, and the data exclusion rates as following (lines 183 – 195):

'Based on the experience from the RIGAB 12 month follow-up, a retention rate of 60% is expected between the first online survey and the third survey wave 9 months later. In addition, the exclusion of further data due to data quality checks was estimated at 25 %. Based on these

response rates, the retention estimates, and the data exclusion estimates, 925 participants need to participate in the first wave to achieve a sample of 370 participants in the third and final wave (60% retention rate, 25% data quality exclusion, 10% buffer). Assuming a linear decline of the retention rate, we estimated that 555 participants have to participate in the second wave (80 % retention rate and 25% data quality exclusion). Therefore, we estimated that we would have to invite 10,153 Tipico account holders in the first wave, of whom about 1,320 would be assumed to respond (response rate 13%) and n = 925 (30% data quality exclusion) could be included in the final sample of the first wave.'

This shows that we also expect data quality to be better in waves 2 and 3 (20% data quality exclusion), as we have only invited participants who have not been excluded due to poor data quality at wave 1.

In addition to these explanations, we would like to address the issue of retention rate in three ways. The first way is also described in the answer to your next point. We have planned very carefully, how to achieve a high retention rate based on previous experience (see the response to your next point for more details). Second, we are right now collecting data for the second wave from which we can conclude that our calculations of the retention rate are reasonable. Specifically, we have already re-contacted 487 participants and 352 have participated in wave 2, which translates to a retention rate of 72.3% (80% was planned). Two more reminder emails will be sent. Due to data quality, 16 % were excluded (25 % were planned). In sum, we are therefore confident that we will achieve the target sample size in wave 3. Third, we will treat the missing values with full information maximum likelihood (FIML). FIML can provide unbiased inferences and valid standard error estimates in the case of nonresponses. We have now included all this information in the manuscript. The first and second points have been included in the 'Recruitment and retention' section (line 220 ff.) and the third point in the 'Statistical analysis' section (line 529).

Minor:

Do the authors have access to gambling records data of the full sample (27,000 account holders) in addition to the 954 who qualified for Wave 1? If so, a more guaranteed way to assess for selection bias would be to conduct basic inferential tests (e.g., median tests) of the important risky behavior measures they have ready access to (e.g., total stake, net losses, variability of betting odds) to see whether/extent to which the two samples differ? This would help to speak to the generalizability of the analytic sample, at least as it relates to the population of Tipico bettors?

Thank you for these important questions. For privacy reasons, we have agreed with Tipico to request double the number of player tracking records so that Tipico will not know who participated in our study. In our first wave we reached n = 954, so we will randomly select another n = 954 out of the 27,000 account holders and request the player tracking data of n = 1908 account holders. Based on your valuable comment, we now plan to compare the gambling behavior data of the participants in our study with the randomly selected account holders to check for selection bias. We have now added this plan in the new section 'Selection bias' (lines 456 – 454) and changed the paragraph in the 'Strength and limitations' section accordingly (lines 565 – 567).

What will the study do for missing cases on net income? Exclude them? If so, why are there no planned procedures to estimate this type of missing data? For example, if the study has access to gambling data, couldn't this be used to estimate net income using multiple imputation procedures?

Thank you for this question, which shows us that we need to clarify our procedures for handling missing data. We expect three different types of missing and excluded data, which we will handle differently:

1. Exclusion due to poor data quality

Participants will be excluded based on our quality controls. These controls include a) the time control (i.e., faster than 10 minutes in the first wave of the survey and faster than 9 minutes in the second and third waves) and an incorrect response to the 'bogus' items (attention control questions such as 'Please select option 4 for this item.'). Participants who fail these quality controls will be completely excluded from the study.

2. Exclusion due to accuracy check

Individual scores of participants in the behavioral tasks will be excluded if the consistency score in the MCQ is below 75% and/or the accuracy in the critical condition of the Go/Nogo task is below chance, which are standard procedures also described in the paper. These missing values will be handled with FIML in the models used to test our hypotheses.

3. Missing values in the player tracking data

It is possible that participants will have some missing data in one of the player tracking data (see new Table 2). To handle missing data, we will use the expectation-maximization (EM) algorithm to estimate the covariance matrix in the factor analysis of the player tracking data (Truxillo, 2005; Weaver & Maxwell, 2014).

In the revised manuscript, we now explain our handling of the missing data in more detail in the subsection 'Missing values, data quality, and data exclusion' (lines 400 – 454).

Review 3

To the authors:

Below are my comments as of Stage 1. I hope you find them constructive.

We would like to thank you for your careful review and constructive feedback. We were pleased to be able to improve our paper in terms of clarity and address other concerns. Please note that the lines referred to in the reply correspond to the manuscript in track change mode (not the original manuscript).

Abstract, math typesetting: Add spaces between equal signs and "n" and other variable letters. n = 370.

It has been corrected.

Introduction, growth of sports betting: I noticed that you're using commas for things like 0-point-3 billion. It should be written as "$0.3 billion in revenue."

It has been corrected.

Introduction, "Researchers and policymakers are concerned...": If I remember correctly, it's the increase in online sports betting _availability_ that people are concerned about. Granted, it's because the increase in availability and increase in legal venues that creates the increase in participation, which might lead to problems at the individual and population levels.

We now explain the argumentation in more detail to prevent misunderstandings (lines 58 – 62):

'Researchers and policymakers are concerned about the potential negative impact of such a large increase in the availability of online sports betting on individual health and public health, as increased availability leads to higher online sports betting participation and cross-sectional studies have linked online gambling participation to a significant rate of gambling problems (Allami et al., 2021).'

Introduction, hypotheses: Here and throughout the report, switch to active voice and take ownership of the methodology. "We proposed three directional hypotheses to test our main assumption."

We agree that the active voice is an important means of emphasizing ownership of the methods. Where appropriate, we have switched to active voice.

Participants, sample size, response rates and dropout rates: You're looking for an ending sample size of n3 = 370, and you expect to keep 40% of those you start with in Wave 1. Assuming that whether one person drops out is independent of whether anyone else drops out, if n1 is the number of people you start with in Wave 1, then n3 has a binomial distribution with n1 trials with probability of success 0.4. Semi-ideal would be a 95% chance that n3 is greater than or equal your target of 370. (n1)(0.4) - 1.645*sqrt(n1*0.4*0.6) = 370. Solve for n1... I'm getting 989. Your calculations have a 0 where I have -1.645, so according to this model, you've got a 50/50 chance of getting the final sample size you want. Your sample size of n1 = 925 might be a bit short, you might get lucky, or you might have a plan to decrease that 60% dropout rate.

Recruitment, actual n1: You have n1 = 954. With that 40% finishing rate, you've got about a three-quarters chance of ending with 370 or more at Wave 3.

Thank you for bringing up this issue. We agree that it is very important to achieve the target sample size. In fact, we did not include random error in our dropout analyses, as you did in your calculation. We would like to point out that we mistakenly referred to the retention rate of 60% in the RIGAB study as the 'dropout rate' in the paper, which we have now corrected. Due to your comment and those of the other reviewers, we realized that we did not explain in enough detail how we calculated the sample size. We have revised the section 'Sample size and power calculations' and gave more detail concerning the estimated response rates, the retention rates, and the data exclusion rates as following (lines 183 – 195):

'Based on the experience from the RIGAB 12 month follow-up, a retention rate of 60% is expected between the first online survey and the third survey wave 9 months later. In addition, the exclusion of further data due to data quality checks was estimated at 25 %. Based on these response rates, the retention estimates, and the data exclusion estimates, 925 participants need to participate in the first wave to achieve a sample of 370 participants in the third and final wave (60% retention rate, 25% data quality exclusion, 10% buffer). Assuming a linear decline of the retention rate, we estimated that 555 participants have to participate in the second wave (80 % retention rate and 25% data quality exclusion). Therefore, we estimated that we would have to invite 10,153 Tipico account holders in the first wave, of whom about 1,320 would be assumed to respond (response rate 13%) and n = 925 (30% data quality exclusion) could be included in the final sample of the first wave.'

This shows that we also expect data quality to be better in waves 2 and 3 (20% data quality exclusion), as we have only invited participants who have not been excluded due to poor data quality at wave 1.

In addition to these explanations, we would like to address the issue of retention rate in three ways. The first way is also described in the answer to your next point. We have planned very carefully, how to achieve a high retention rate based on previous experience (see the response to your next point for more details). Second, we are right now collecting data for the second wave from which we can conclude that our calculations of the retention rate are reasonable. Specifically, we have already re-contacted 487 participants and 352 have participated in wave 2, which translates to a retention rate of 72.3% (80% was planned). Two more reminder emails will be sent. Due to data quality, 16 % were excluded (25 % were planned). In sum, we are therefore confident that we will achieve the target sample size in wave 3. Third, we will treat the missing values with full information maximum likelihood (FIML). FIML can provide

unbiased inferences and valid standard error estimates in the case of nonresponses. We have now included all this information in the manuscript. The first and second points have been included in the 'Recruitment and retention' section (line 220 ff.) and the third point in the 'Statistical analysis' section (line 529).

Measures, factor analysis of gambling measures: What threshold will you use to determine how many factors you will use in your models? Just the eigenvalues? What will you do if the first factor explains, for example, only 10% of the variation between gamblers?

We now describe more transparently how we will select the number of factors to operationalize risky gambling behavior. Factors will be selected using the eigenvalue, scree plot, parallel analysis, and interpretability. A detailed description of our selection criteria has been added to the subsection 'Operationalization and study materials' (lines 351 – 366). Based on our experience with the player tracking data in our previous study, we are confident that the first factor will explain more than 10% of the variance. In a factor analyses with our player tracking data from the RIGAB study, the first factor explains 33% of the variance.

Statistical analysis, factor analysis: Consider including a rundown of the factors and which gambling measures are most salient. You might be able to add descriptive labels to the factors. You might be able to separate the gamblers into different groups based on whether they're extremely high in one factor or extremely low in another. For example, one factor might heavily lean on amount wagered on in-game bets, and there might be people who have an extremely high value for that factor (in-game betting specialists?) and people who have a value closer to zero (pure pre-game bettors?).

Thank you for sharing this idea with us. This is also an aspect we would be interested in. In a recent paper based on the RIGAB data (Czernecka et al., in prep.), we have found that most of the variables in the gambler tracking data can significantly discriminate individuals with a gambling disorder from those without. In line with previous findings (Braverman & Shaffer, 2012), we assume to find factors related to gambling intensity and variability. Both factor scores would certainly be elevated in people with a gambling disorder. However, our assumptions are not specific enough and we consider this approach rather exploratory. We therefore decided to document your interesting suggestion and include it in further exploratory analyses after testing our confirmatory hypotheses.