PLOS ONE

The Ordinal Effects of Ostracism: A Meta-Analysis of 120 Cyberball Studies

--Manuscript Draft--

Manuscript Number:	PONE-D-15-02806R1
Article Type:	Research Article
Full Title:	The Ordinal Effects of Ostracism: A Meta-Analysis of 120 Cyberball Studies
Short Title:	ORDINAL EFFECTS OF OSTRACISM
Corresponding Author:	Chris H.J. Hartgerink Tilburg University Tilburg, NETHERLANDS
Keywords:	Cyberball; meta-analysis; ordinal; ostracism
Abstract:	We examined 120 Cyberball studies (N = 11,869) to determine the effect size of ostracism and conditions under which the effect may be reversed, eliminated, or small. Our analyses showed that (1) the average ostracism effect is large (d > |1.4|) and (2) generalizes across structural aspects (number of players, ostracism duration, number of tosses, type of needs scale), sampling aspects (gender, age, country), and types of dependent measure (interpersonal, intrapersonal, fundamental needs). Further, we test Williams's (2009) proposition that the immediate impact of ostracism is resistant to moderation, but that moderation is more likely to be observed in delayed measures. Our findings suggest that (3) both first and last measures are susceptible to moderation, and (4) time passed since being ostracized does not predict effect sizes of the last measure. Thus, support for this proposition is tenuous, and we suggest modifications to the temporal need-threat model of ostracism.
Order of Authors:	Chris H.J. Hartgerink
	Ilja van Beest
	Jelte M. Wicherts
	Kipling D. Williams
Opposed Reviewers:	
Response to Reviewers:	Ilja van Beest Faculty of Social and Behavioral Sciences Tilburg University P.O. Box 90153 5000 LE Tilburg The Netherlands Prof. Dr. Nico van Yperen Academic Editor, PLOS ONE Rijksuniversiteit Groningen Manuscript ID PONE-D-15-02806 Dear Prof Dr. van Yperen, Thank you for your kind words about our work and for the opportunity to submit a revised version of our paper “The Ordinal Effects of Ostracism: A Meta-Analysis of 120 Cyberball Studies" to PLOS ONE. As you recommended, we have carefully considered each of the comments made by the reviewers, paying special attention to those highlighted by you in your letter. A detailed overview of our revisions is included below. For your convenience we copied the three reviews and added a detailed description of how we made the appropriate changes immediately below each comment. We believe that the changes we made have substantially improved the manuscript and made our contribution stronger. We warmly thank you for your help in achieving this and look forward to your final decision.

Kind regards, also on behalf of Chris Hartgerink, Jelte Wicherts, and Kip Williams

Ilja van Beest

***

Reviewer #1: The authors have conducted a meta-analysis of studies using the Cyberball game, which manipulates the degree of social inclusion versus ostracism experienced by participants. Particular focus in this meta-analysis is on the immediate and delayed effects of the experimental manipulation and on examining whether immediate or delayed effects are more susceptible to the moderating influence of other factors.

In general, I think that PLoS One is an appropriate outlet for this meta-analysis and I would support its publication. However, below I list a number of general issues, concerns, comments, and appeals for clarification that I think the authors need to address first.

General Issues / Concerns / Comments:

#1

page 7: Author predictions were used to determine how an interaction should be coded. Was there always a clear prediction given by authors so that this decision could be made unambiguously? If not, was the intercoder reliability of these assessments measured? Aside from this, we know that some 'predictions' are actually generated post-hoc, after the results have become available. That is a limitation that should be acknowledged.

Answer

Of the 120 studies that were investigated, 52 studies contained an interaction. The prediction in these 52 studies, was based on the explicit prediction of the authors of the manuscript. Moreover, the first authors (Chris and Ilja) checked and discussed each paper until consensus was reached. We did not record these discussions and intercoder reliability cannot be assessed. We did provide a case by case description of all studys on OSF.

We acknowledge that the predictions of the primary studies could be post-hoc and this is now acknowledged in the revised manuscript. We now say

A potential limitation of our decision to follow the prediction of the authors is that the predictions may have been generated post-hoc on the basis of observed outcomes.

on page 7 line 135.

#2

page 10, line 208: In studies with more than one additional factor (besides the ostracism factor), the authors "collapsed effect sizes across the factor that authors expressed least interest in." I can imagine that this decision cannot always be made with 100% certainty. Did the authors attempt to estimate the intercoder reliability of these assessments?

Answer

Seventeen of the 52 studies with a cross-cutting variable involved designs that were more complex than the 2x2 design. In these studies, the selection decisions were jointly made by Chris Hartgerink and Ilja van Beest. Intercoder reliability was not assessed.

#3

page 10, line 224: I know from personal experience that one of the last things that authors of a meta-analysis want to hear is: Your search is outdated. Indeed, a meta-analysis may go through through several (re)submission rounds before being accepted/published and the date of the search then increasingly falls further behind. There is in principle no need to demand an update, so I will not insist. However, are the authors aware of any additional studies that have become available after their search was concluded?

Answer

We agree that the search for additional studies is time-consuming and that one should always chose a moment to stop updating the database. Nevertheless, we conducted another search in Web of Knowledge for Cyberball studies, which resulted in 71 hits for 2013-2015 (searched on March 17, 2015). After inspecting which of these studies would have met our inclusion criteria, 29 remained after our previous end date (April 2013). These 29 references are available here (EndNote format). Of these 29, we already included 2 studies that were not published when we collected them, and 14 contained a cross-cutting variable. Given the current size of the database and the sample sizes in these new studies, we do not expect them to significantly change any of our core conclusions. Hence, we decided not to redo all of the analyses using this updated database.

#4
page 13: I am wondering about the selection/coding of the first and last measure. Was there never any ambiguity regarding the order in which instruments were administered? Also, if authors said that they used measures X, Z, and Z after the game, the actual order may have been different.

Answer
We based the coding of the first- and last measure on the information presented in the paper describing the primary study. This information was straightforward and we did not encounter ambiguity regarding the order in which the instruments were administered. We acknowledge that people may have included more measures than reported and that unreported measures remain unaccounted for, such that the estimate for time between the first and last is a crude one. In other words, we could not get better information than that reported in the paper, which is why we retain the information reported in the paper as the most viable situation.

#5
page 13: First and last measures were classified into four categories (interpersonal, intrapersonal, fundamental needs, or model correspondence). So, if I understand the authors correctly, first a measure was chosen as being first/last and then this classification was made (so there is always exactly one first measure and if the study applied multiple/delayed assessments, there is always exactly one last measure). Can the authors please confirm/clarify this?

Answer
We hereby confirm that every study contained a first measure and if present, a last measure. Table 2 illustrates this, where some studies do not contain an effect on the last measure.

#6
page 13: Also, does that imply that the first measure may have assessed, for example, intrapersonal effects, while the last measure may have assessed, for example, interpersonal effects? Or in other words, is it possible that the effect size estimates in Table 2 (d_T1 and d_T2, and similarly, Delta-d_T1 and Delta-d_T2) actually reflect different measurement types? This needs to be clarified, since this has major implications for the interpretation of the results reported on pages 25 to 27.

Answer
Yes, this is correct. Figure 2 separates the effects per type of measure and shows that results are consistent across the different types of dependent variables, except for interpersonal behavior (as mentioned in the text).

#7
page 18, line 350: I am not sure if "standardized simple effects across the ostracism factor" is appropriate terminology here (and elsewhere in the paper). In a two-way factorial design, a "simple effect" is the effect of one factor *within* one of the levels of the other factor. So, if that other factor has two levels, then there would be two simple effects. That would apply to each time point, so in a 2x2 design with multiple measures (one of which is the first and one is the last measure), there would be 4 (not 2) simple ostracism effects. However, if I understand the authors correctly, they are not computing simple effects here, but marginal/main effects for the first and for the last measure (i.e., the difference between the ostracism and inclusion levels averaged over

any other factors). Please clarify this (and the terminology throughout the manuscript).

Answer
We did intend simple effects, as we calculated four simple effects for the ostracism factor (one in the moderated conditions, one in the non-moderated conditions, for both first and last measure). The reviewer refers to the set of 52 studies where a second factor is included, where we calculated the simple effect of ostracism within the non-moderated level. We clarified this in the revised manuscript. Specifically, we now write:
Standardized effects were calculated across the ostracism factor, where the 52 studies with a cross-cutting variable were included as a simple effect of ostracism within the non-moderated level.
On page 18 ~ line 349. Additionally, we deleted the following to prevent confusion (lines 355-356):
Non-factorial studies delivered only simple effects for the first and last measure, and no interactions

#8
page 18: The description of the interaction effect given here (and on the previous pages and also the appendix) suggests that moderators of the ostracism effect can take on only two values/levels. However, was that always the case?

Answer
Moderator factors could include more levels, in which case we selected the two conditions that were the farthest apart in design. For example, if a study included an ostracism factor (included or ostracized) and a players factor (3, 5, 10, 15 players) as a moderator, we used the 3 and 15 player levels. Selection based on the factorial levels occurred in 10 studies. We mention this number in the text of the revised manuscript (page 18 line 359)

Table 2:

#9
1) I see many rows where "First author" and "Year" is identical. Can the authors explain how this arises?

Answer
We thank the reviewer for this comment. The reason is that papers may contain multiple studies. To clarify this, we now added a note.
Multiple rows for the same first author and year is possible due to multiple studies across papers.

#10
2) In the table notes, the authors write: "Non-integer Ns arise from division of full sample N for included conditions, appropriate due to random assignment." I don't understand what the authors mean by this (and I could find no further discussion of this in the paper).

Answer
Ns of for example 12.333 arise from a 3-condition design, where random assignment was used. If N per condition was not given, we divide total N (e.g., 37) by the number of conditions (3) to come to a condition N estimate. To clarify we added an example in the table note:
(e.g., two conditions out of 3, when sample is 56: (56 / 3) × 2 = 37.333)

#11
3) It appears that multiple estimates are often obtained from the same study. Given that "N" differs for these rows, these effects seem to be based on different samples, so within a particular study, the estimates may be independent. However, that still does not preclude the possibility that multiple estimates obtained from the same study are more similar to each other than estimates obtained from different studies. In other words, the data seem to have a multilevel structure, which would imply the need to employ an appropriate multilevel meta-analysis model that accounts for such dependencies (e.g., by adding a random effect at the study level to the current model).

Answers
The reviewer notes that the data may be interdependent within an analysis; this is incorrect. Effects that go into the same meta-analysis are independent (i.e., one effect per study): every row is an independent study, which also explains the difference in N. However, the reviewer is correct in stating that from one paper multiple independent studies can be included. This multilevel modeling is therefore not necessary.

#12
page 25: I assume the authors applied the version of Egger's regression test that relates the effect size estimates to their standard errors. For standardized mean differences, the standard error depends on the size of the effect, which can cause spurious associations especially when effects are large. Similar deficiencies of the test have been observed when using effect size measures based on dichotomous data (e.g., risk/odds ratios or risk differences). For a more appropriate version of the test, the authors should use some measure of precision that does not depend on the size of the effect, the obvious choices being the sample size, the inverse sample size, or square-root transformations thereof.

Answer
As requested by the reviewer, we conducted these regression tests with 1/N as predictor. Results are the same as the Egger’s test with standard error as predictor and is therefore not adjusted further in the manuscript. We include a footnote in the methods section of the manuscript that reads:
Due to the dependency between the standardized effect size and the standard error, we also ran an alternative version of the Egger’s test that regresses on 1/N. These analyses yielded highly similar results.

#13
page 25: Coding the estimated time between exclusion and the moment at which the last measure was taken in *seconds* seems artificially precise. Did the authors calculate the intercoder reliability for these estimates based on independent coders? Also, please rescale this moderator into some larger units (e.g., minutes) which avoids the extremely small coefficient (.0001). In addition, since this is one of the primary hypotheses tested in the paper, please provide a scatterplot of the time variable against the effect size estimates.

Answer
Following the suggestion of the reviewer we rescaled the time estimate into minutes. The results have been adjusted accordingly.
Also note that the time estimation was based on the number of items times the six second rule, plus any additional time mentioned in the paper. This information was readily available in all manuscripts although we acknowledge that it is possible that not all dependent variables were disclosed in a paper describing the study (see also our answer reviewer 1, #4). As mentioned, in the 68 studies without cross-cutting variable were coded by Chris Hartgerink, the 52 with a cross-cutting variable were coded by both Chris Hartgerink and Ilja van Beest. Consensus was readily reached and we did not collect quantitative information to calculate intercoder reliability.
Following the suggestion of the reviewer, we now provide scatterplots of time versus effect (simple and interaction on timepoint two) in the Supplemental Materials of the revised manuscript.

#14
page 27: Same issues apply here. I cannot imagine that two independent coders would ever come to the exact same assessment when coding time in seconds. Also, please rescale time to avoid the overly small coefficient. And please provide a scatterplot.

Answer
See reviewer #1, answer #13.

#15
page 28 and Figure 2: As far as I can tell, here the authors are indeed talking about simple effects (e.g., "the between-subjects effect of being ostracized with no moderator present, whereas moderated ostracism effect refers to being ostracized with a moderator present"). Earlier, the authors also talked about "simple effects" (which I

think are actually main effects -- see my earlier comment -- but maybe I am misunderstanding what the authors did). Please clarify this.

Answer
See reviewer #1 answer #7.

#16
Also, if I understand Figure 2 correctly, I would assume then that the *difference* between, let's say, the points for "All" in panels (1) and (2) is equal to the *difference* between the points for "All" in panels (5) and (6) (since the difference between the two simple effects for factor A within the two levels of factor B must be equal to the difference between the two simple effects for factor B within the two levels of factor A). However, visual inspection suggests that this may not the case. Can the authors clarify?

Answer
We are not sure whether we understand the question. It seems that the reviewer postulates that the difference in the simple effects for ostracism on the different moderator levels is supposed to be equal to the difference in simple effects for the moderator levels on the ostracism levels. Below we provide an example that this would be incorrect and that simple effects do differ.
N-modmod
Ostr57
Incl23

In this case, the simple effect of ostracism is 5-2 = 3 for the non-moderator level and 7-3 = 4 for the moderated level. For the simple effect of moderator within the ostracism level, 5-7 = -2 and within the included level 2-3 = -1. Correspondingly, simple effects all differ and are not required to be equal, as the reviewer proposes.

#17
page 30, line 514: "Model indicates" -- which model?

Answer
The model pertained to a subset included throughout the analyses. To avoid confusion we rewrote the note under table 3 to read similar to Figure 2
The subset labeled “All” contains all measures. The subset labeled “Fundamental” contains only fundamental need measures. The subset labeled “Intrapersonal” contains all intrapersonal measures. The subset labeled “Interpersonal” contains all interpersonal measures. The subset labeled “Model” contains those where first measures is immediate and last measure is delayed. See Supplement S4.
On page 28 this was clarified under the heading Measures, where the subsets are named.

#18
page 30, lines 515-516: I don't understand what the authors mean by "listwise deletion for equal ks across time points". Please clarify.

Answer
To clarify what we mean by listwise deletion we adjusted the sentence as follows:
Listwise deletion ensures that estimates are made on full rows in the data. Listwise deletion was applied in all the subsets, which only altered results for interpersonal measures.

#19
page 30, line 520: What estimates did the authors use for these analyses? The estimates shown in Table 2 or the "simple effects" that went into the analyses that led to Figure 2? I assume the former values were used, but please clarify this. Also, if my assumption is correct, then as far as I can tell, listwise deletion (due to incomplete information about the predictor variables) led to the removal of 120 - 45 = 75 estimates for T1 and 95 - 41 = 54 estimates for T2. Is that correct? If so, then this should be mentioned as a limitation.

Answer

The analyses were based on the ostracism effect across all 120 studies (as in Table 2 column d T1). However, due to listwise deletion the number of effects indeed reduced the number of effects included and now reads:
To inspect for structural and sampling effects of the studies, we ran mixed-effect models on the 120 ostracism effects, on both the first and the last measure. Due to listwise deletion, only 45 of 120 effect sizes remained on the first measure and 41 of 95 effect sizes for the last measure.

#20
pages 30, line 527: The dfs for the Q_E-test are 32. With k = 45, this implies that the model must have contained 45 - 32 = 13 fixed effects (including the intercept). However, in Table 4, I only count 12 coefficients.

Answer
We thank the reviewer for noting this error. The dfs should indeed be 33. This is now adjusted in the revised manuscript.

#21
page 31, line 537: The dfs for the Q_M-test are 12. Assuming that the intercept was not part of the coefficients tested, this implies that the model included 13 fixed effects. However, I only count 12 coefficients in Table 5.

Answer:
We again thank the reviewer for noting this error. The df should be 11 and is adjusted in the revised manuscript.

#22
page 31: Please report the results from the Q_E-test here as well.

Answer:
We added the results. On page 32 of the revised manuscript we now say:
QE (29) = 214.69, p < .0001

#23
Tables 4 and 5: For a categorical predictor with more than 2 levels, please provide a test of the factor as a whole (i.e., an omnibus test of the coefficients corresponding to the factor). Also, the tables only show the results of tests comparing each level against the reference level, but there may be significant differences when comparing other levels against each other. Please examine/report this.

Answer:
The Q_M test is an omnibus test and is reported. The dummies are indeed only compared to the reference group. Moreover, we already included confidence intervals in the original version of our manuscript. These CIs indicate that all comparisons between these dummies will yield similar results (overlapping CIs). Indeed, the requested analyses confirmed this:
If we only look at the countries, QM(df = 2) = 0.3494, p-val = 0.8397, first measure, QM(df = 2) = 2.6394, p-val = 0.2672, last measure.
If we only look at the different needs scales, QM(df = 4) = 6.0702, p-val = 0.1940, first measure, QM(df = 4) = 0.4257, p-val = 0.9803, last measure.
Because these analyses provide the same information as the overlapping confidence intervals we decided not to incorporate them in the revised manuscript.

#24
page 41, line 738: I don't understand what the authors mean by "difference index" or how this was coded. What "value" are the authors referring to when they write: "coded value on first measure minus coded value on last measure"? In fact, I have a hard time understanding this entire paragraph.

Answer
We thank the reviewer for this comment. We wanted to explain that differences in findings between first and last measurement could not be attributed to differences in types of dependent variables. We now write (on page 41-42):
Importantly, we did observe that the confidence intervals of both the first and last

measure did not overlap, suggesting that there is a difference in effect size between first and last measure. The question then is whether this difference is indeed caused by time of measurement or in part caused by the type of measurement used across the two different time points. This explanation can be addressed by inspecting whether the composition of measures is different across time points. On the first measure 0.84 was intrapersonal self-report, 0.02 was intrapersonal physiological, 0.01 was intrapersonal other, 0.08 was interpersonal anti-social, 0.03 was interpersonal pro-social, and 0.01 interpersonal other. On the last measure 0.79 was intrapersonal self-report, 0.04 was intrapersonal physiological, 0.02 was intrapersonal other, 0.05 was interpersonal anti-social, 0.08 was interpersonal pro-social, and 0.01 was interpersonal other. This shows that the different types of dependent variables are similarly distributed across time points (maximum discrepancy of 4.9 percentage points). Substantive differences in proportions of measures across time points are minimal and thus form an unlikely driving force for our findings.

Minor Issues:

#25
Maybe this term is well understood by the intended target audience, but I find the term "cross-cutting variable" less than clear. Why not just call them "other factors" or something along those lines?

Answer
The term cross-cutting factor is a standard term in the Cyberball field. It refers to design in which the ostracism manipulation (inclusion vs ostracism) is orthogonally crossed with another manipulation (e.g., ingroup vs outgroup). Additionally, because we also include other moderator variables (i.e., time, structural, sampling), we use “cross-cutting” as a term to prevent confusion. Cross-cutting refers to the 52 studies that explicitly manipulated a factor in the experimental design. The other moderator variables (e.g, time, structural, sampling) were investigated for all 120 studies.

#26
page 3, line 47: The "(4)" is superfluous (or also number the other moderator types).

Answer
Adjusted

#27
page 3, line 53: Write out "i.e." when used outside of parentheses.

Answer
Adjusted (also checked rest of i.e. occurrences)

#28
page 3, line 54: "an unique" should be "a unique" (the use of "a/an" is not based on the spelling of the first letter of the following word, but its pronunciation).

Answer
Adjusted

#29
page 7, line 150: "set-up" should be "set up" (set-up or setup is a noun).

Answer
Adjusted

#30
page 9, line 182: "extend" should be "extent" (the latter is the noun). And the more common phrasing would be "to a large extent".

Answer
Adjusted

#31

page 11, line 226: Write out the acronym (SPSP) the first time it is used.

Answer
Adjusted

#32
page 13, lines 291 and 293: Since you are giving examples here ("e.g.,"), the "etc." at the end is superfluous.

Answer
Adjusted

#33
page 14, line 301: Missing comma after "e.g.".

Answer
Adjusted (checked all occurences of e.g.)

#34
Table 1, table notes: I think the "whereas column wise" should be "whereas row wise".

Answer
Adjusted

#35
page 41, line 754: "conditional on that these measures are valid" is very odd phrasing.

Answer
Deleted this sentence.

#36
The Oxford comma is used inconsistently throughout the manuscript.

Answer
We checked the manuscript for consistency and adjusted where needed.

Appendix:

#37
1) df_w needs to be defined.

Answer
Adjusted. Added that this is equal to conditions minus 1.

#38
2) The top part of a fraction is called "numerator", not "nominator".

Answer
Adjusted

#39
3) Isn't the first term in the numerator the ostracism effect *in the non-moderated/control condition* (and the second term is the effect in the moderated condition)?

Answer
We calculated it in the order we describe. It can also be done the other way around, which would lead to a change in interpretation but equal results.

#40
4) In what sense does Delta-d "correspond" to partial eta-squared of the interaction? Numerically it cannot be the same (partial eta-squared must be between 0 and 1, while Delta-d as defined is not a proportion and may be larger than 1 and can be negative).

Answer
When the resulting d is transformed into a squared correlation coefficient it gives the exact same value. This is highlighted in the Appendix and now reads
When transformed to a squared correlation coefficient, this Δd corresponds to the partial eta-squared of the interaction.

#41
5) Please add ^2 to s_g and s_d to make it clearer that these are variances.

Answer
Done.

#42
Final comment: In the spirit of open science, I appreciate the use of OSF and the authors' transparency in conducting this meta-analysis.

Answer
Thank you. We also like to thank the reviewer for the thorough review and thus for making this a better manuscript.

Reviewer #2:
#1
Overall this study looks competently executed and acceptable for publication. My only real concern is that authors could have done more to explore and account for the variability in their data. The meta-analysis demonstrates that the variability was considerable, but beyond establishing that moderators exist, the researchers appear to be not overly concerned with the question what is causing this variation. That leaves me slightly unsatisfied at the end: all this effort to conduct a meta-analysis, and the main thing we learn is that (a) the effect of rejection is strong (something we knew because it has been shown time and again), (b) the first sharp shock diminishes over time (new to me, but then I'm not an expert), and (c) the intensity of that shock depends… If authors were willing to stick their finger out a bit more and clarify just what this depends on, I'm sure I would find the study more valuable than it is now. I don't care if their hypotheses were deposited beforehand: exploring is a scientists' duty, as much as hypothesizing in advance (e.g., Tukey). But to be clear: this is just meant an encouragement; it's very much up to to the authors to decide what course of action to pursue.

Answer
We thank the reviewer for his/her kind words and regarding the manuscript as competent and acceptable for publication. We agree with Reviewer #2 that exploring the data is a valuable avenue for any study, including this meta-analysis. As a matter of fact, we were also puzzled by the heterogeneity in the data and we therefore conducted several exploratory analyses to understand this heterogeneity. The most important exploratory analysis that we conducted was the one in which we selected the most homogenous subset possible (i.e., only immediate fundamental need measures, 30 throws, 3 players), but still found high heterogeneity. Meta-regressions also failed to indicate any explanation for the heterogeneity. We agree that further exploration is definitely interesting, but also believe that we exhausted all possibilities that were available to us in the current dataset.

Some other points that would help authors improve the paper up to a level that would match my expectations for PLOS One standard mainly concern the quality of the writing and the care about the argument being made. The introduction reveals that authors could have spent some more care writing (and perhaps thinking about) their subject. Suffice to say that it's important to be precise. Some examples:

#2
"Cyberball participants simply do not obtain a ball and thus need to infer that they are excluded" I think authors are trying to say something about implicit and explicit exclusion here. I also think they are trying to say something about acting together versus communicating with each other. But it's not being said.

Answer

This sentence was deleted, because the preceding sentence already contains the information.

#3
The sentence “This focus on ostracism makes it an unique paradigm...” is clearly erroneous, because it is not the focus on ostracism that makes cyberball unique.

Answer
The first paragraph in the Historical background section is changed into:
Cyberball was introduced in 2000 as a means to study ostracism, that is: being excluded and ignored [1]. This focus of Cyberball on ostracism sets it apart from other paradigms that are tailored to study rejection, such as the future life rejection [2], the get-acquainted paradigm [3], and the autobiographical memory manipulation (i.e., remember a time when you were excluded [4]). The difference is that participants in Cyberball are not explicitly informed that they are excluded whereas in the other paradigms participants are provided a reason pertaining to why they are excluded.

#4
Further on, a sentence such as “research suggests that most people are ignored and excluded at least once a day” sits happily side by side with the sentence “research on school shootings has suggested a direct link between ostracism and revenge”. This could be spelled out more clearly. If everyone is a victim of exclusion, then obviously those who go on a shooting spree are, too. So is the point that ostracism is a frequently occurring post-hoc justificationfor this kind of behavior?

Answer
We adjusted the sentence. It now reads:
The social relevance is further evident in that ostracism not only affects the person who is ostracized (intrapersonal effects), but often also others (interpersonal effects). As a grim example, research on school shootings has suggested a direct link between ostracism and revenge. People who were ostracized may retaliate by murdering those responsible and sometimes even innocent bystanders [5].
#5
Further on authors write “This initial response is theorized to be socially painful, threatening [9] and easily detectable due to evolutionary over-sensitivity to cues of ostracism [12].” In a sentence such as this, please carefully distinguish phenomenon and hypothesis. There is abundant evidence for the first inference, but the evolutionary origins of this phenomenon can only be inferred indirectly from its existence and prevalence.

Answer
We adjusted the sentence. It now reads:
This initial response is theorized to be socially painful, threatening [9] and, following overdetection theory [12], should be easily detectable due to evolutionary over-sensitivity to cues of ostracism.

#6
It is stated that all selections and hypotheses were preregistered on OSF. But what is not spelled out is whether authors tried to learn something new from their data by exploring it?

Answer
We explored several avenues. For example see reviewer #2,
answer #1, but also
answer below (reviewer #2,
answer #7)

#7
“Examples of interpersonal measures are donations to charity, helping behavior, money allocations in economic games, and aggression measures such as irritating sounds blasts or hot sauce allocation.” Please split the effects of positive and negative behaviors—they are qualitatively too distinct to be lumped together in this way. Later on I noted that K=10 for these studies (?). If small K was the reason for lumping things

together please explain the criteria and total K in this section to help readers understand your decision making process.

Answer
These were indeed split into positive (pro-social) and negative (anti-social) behaviors initially and were indeed lumped together due to small K, hence, low power for detecting moderation effects. For the first measure, there were 14 interpersonal measures, of which 4 are positive and 10 negative. For the last measure, there were 14 interpersonal measures, of which 8 are positive and 6 negative. We added a sentence in the manuscript to clarify this. One page 8 of the revised manuscript we now say:
These were initially coded into pro- and anti-social, but were collated into the category interpersonal due to small k the first measure (4 and 10, respectively) and last measure (8 and 6, respectively).

#8
For various decisions to include or exclude studies or factors, please provide an indication of the number of studies affected by your decision. E.g., "continuous variables that were dichotomized into factorial levels were also collapsed due to the many problems dichotomization can cause". How many studies were collapsed in this way? I'm trying to assess the impact of your coding decisions.

Answer
This collapsing occurred a total of four times, for the studies from (i) Stock 2011, (ii) two studies from Boyes 2009, and (iii) Zadro 2006. We added this number in the manuscript on page 10..
Some other minor points:

#9
"we used the metafor package": include version.

Answer
Version 1.9-5. Added in the manuscript.

#10
I do not understand this sentence: "Model indicates that the first measure was indeed reflexive and the last measure reflective."

Answer
The model pertained to a subset included throughout the analyses. To avoid confusion we rewrote the note under table 3 to read similar to Figure 2
The subset labeled "All" contains all measures. The subset labeled "Fundamental" contains only fundamental need measures. The subset labeled "Intrapersonal" contains all intrapersonal measures. The subset labeled "Interpersonal" contains all interpersonal measures. The subset labeled "Model" contains those where first measures is immediate and last measure is delayed. See Supplement S4.

#11
"meta-analyses" is plural

Answer
Adjusted

#12
"by a large extend"
= to a large extent

Answer
Adjusted

Reviewer #3: This study is a system review and meta-analysis of cyberball studies for effect size of ostracism. The manuscript is well-written and provides many detailed information for readers. The statistical analysis is rigorous and well-thought. The primary and secondary hypotheses are clearly stated. The results and discussion are

	also clearly presented. I have following comments. We thank the reviewer for his kind words and stating that our analyses are rigorous and the manuscript is well-written. #1 1. First, I appreciate the authors’ efforts in providing detailed information about the data and implementation, which greatly improve the transparency and reproducibility of the research. More importantly, the information is very helpful for readers to have an objective view of this study. Answer Thank you for your kind words. #2 2. I would suggest moving the “code procedure” sub-section in Method section to supplementary. Although the code procedure is very important and helpful for some readers, it is too technical for most of readers. Answer Although we understand the concerns for the technicalities, the supplement is meant for additional information only, while we consider the coding a crucial aspect of our method. We had thorough discussions on whether it was possible to have directional coding in spite of the bidirectionality of the expected effects and we think a reader will want to know how we were able to make directional claims despite this variety of measures and predictions. Hence, we think it is vital to retain this in the main manuscript. #3 3. I suggest adding a figure for study inclusion criteria. Many system review and meta-analysis paper in PLoS ONE use a figure to demonstrate the procedure for selecting studies. Answer The manuscript contains the PRISMA flowchart in the supplemental materials that addresses this point. We added the flowchart in the manuscript. #4 4. It’s better to present the information in Table 2 as a forest plot, while putting the table 2 in supplementary. A forest plot summarizes the information and gives readers a intuitive understanding. Answer We agree that a forest plot gives an intuitive overview of the effects. However, we think that the forest plot across 120 effects will be too sizable. More importantly, the American Psychological Association prescribes that meta-analyses are to report the data on which main analyses are performed in a table. We therefore think it is more informative to retain the current format.
Additional Information:	
Question	**Response**
Financial Disclosure Please describe all sources of funding that have supported your work. A complete funding statement should do the following: Include **grant numbers and the URLs** of any funder's website. Use the full name, not acronyms, of funding institutions, and use initials to identify authors who	The preparation of this article was supported by grant number 016-125-385 from the Netherlands Organization for Scientific Research (http://nwo.nl) awarded to Jelte M. Wicherts and by the National Science Foundation (http://nsf.gov) under Grant #BCS-1339160 awarded to Kipling D. Williams.

received the funding. **Describe the role** of any sponsors or funders in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. If they had no role in any of the above, include this sentence at the end of your statement: "*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*" If the study was **unfunded**, provide a statement that clearly indicates this, for example: "*The author(s) received no specific funding for this work.*" * typeset	
Competing Interests You are responsible for recognizing and disclosing on behalf of all authors any competing interest that could be perceived to bias their work, acknowledging all financial support and any other relevant financial or non-financial competing interests. Do any authors of this manuscript have competing interests (as described in the PLOS Policy on Declaration and Evaluation of Competing Interests)? **If yes**, please provide details about any and all competing interests in the box below. Your response should begin with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests:* **If no** authors have any competing interests to declare, please enter this statement in the box: "*The authors have declared that no competing interests exist.*" * typeset	The authors have declared that no competing interests exist.
Ethics Statement You must provide an ethics statement if	N/A

your study involved human participants, specimens or tissue samples, or vertebrate animals, embryos or tissues. All information entered here should **also be included in the Methods section** of your manuscript. Please write "N/A" if your study does not require an ethics statement.

Human Subject Research (involved human participants and/or tissue)

All research involving human participants must have been approved by the authors' Institutional Review Board (IRB) or an equivalent committee, and all clinical investigation must have been conducted according to the principles expressed in the Declaration of Helsinki. Informed consent, written or oral, should also have been obtained from the participants. If no consent was given, the reason must be explained (e.g. the data were analyzed anonymously) and reported. The form of consent (written/oral), or reason for lack of consent, should be indicated in the Methods section of your manuscript.

Please enter the name of the IRB or Ethics Committee that approved this study in the space below. Include the approval number and/or a statement indicating approval of this research.

Animal Research (involved vertebrate animals, embryos or tissues)

All animal work must have been conducted according to relevant national and international guidelines. If your study involved non-human primates, you must provide details regarding animal welfare and steps taken to ameliorate suffering; this is in accordance with the recommendations of the Weatherall report, "The use of non-human primates in research." The relevant guidelines followed and the committee that approved the study should be identified in the ethics statement.

If anesthesia, euthanasia or any kind of animal sacrifice is part of the study, please include briefly in your statement which substances and/or methods were applied.

Please enter the name of your Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board, and indicate whether they approved this research or granted a formal waiver of ethical approval. Also include an approval number if one was obtained. **Field Permit** Please indicate the name of the institution or the relevant body that granted permission.	
Data Availability PLOS journals require authors to make all data underlying the findings described in their manuscript fully available, without restriction and from the time of publication, with only rare exceptions to address legal and ethical concerns (see the PLOS Data Policy and FAQ for further details). When submitting a manuscript, authors must provide a Data Availability Statement that describes where the data underlying their manuscript can be found. Your answers to the following constitute your statement about data availability and will be included with the article in the event of publication. **Please note that simply stating 'data available on request from the author' is not acceptable.** ***If, however, your data are only available upon request from the author(s), you must answer "No" to the first question below, and explain your exceptional situation in the text box provided.*** Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	Yes - all data are fully available without restriction
Please describe where your data may be found, writing in full sentences. **Your answers should be entered into the box below and will be published in the form you provide them, if your manuscript is accepted.** If you are copying our sample text below, please ensure you replace any instances of XXX with the appropriate details. If your data are all contained within the	All relevant data are within the paper and its Supporting Information files, and are also accessible via https://osf.io/ht25n/.

paper and/or Supporting Information files, please state this in your answer below. For example, “All relevant data are within the paper and its Supporting Information files.” If your data are held or will be held in a public repository, include URLs, accession numbers or DOIs. For example, “All XXX files are available from the XXX database (accession number(s) XXX, XXX)." If this information will only be available after acceptance, please indicate this by ticking the box below. If neither of these applies but you are able to provide details of access elsewhere, with or without limitations, please do so in the box below. For example: “Data are available from the XXX Institutional Data Access / Ethics Committee for researchers who meet the criteria for access to confidential data.” “Data are from the XXX study whose authors may be contacted at XXX.” * typeset	
Additional data availability information:	

January 20, 2015

Dear PLOS staff,

We hereby submit our manuscript ‘Ordinal Effects of Ostracism: A Meta-Analysis of 120 Cyberball Studies’. We would appreciate it if you could consider our work for publication in the *PLOS: ONE* journal. This is an original manuscript, and is not under consideration elsewhere. The main text of the manuscript is 11,345 words long, and is accompanied by 2 figures and 5 tables. Chris H.J. Hartgerink and Ilja van Beest contributed equally to this work and share first authorship. Correspondence concerning this article should be addressed to Ilja van Beest, i.vanbeest@uvt.nl

Cyberball was introduced by JPSP in the year 2000 as a new method to study ostracism and has now been cited 898 times according to google scholar (Williams, Cheung, & Choi, 2000). In the current manuscript we present a meta-analysis of all the published and unpublished studies that have been conducted with Cyberball since. We expect our analyses to spur debate, as it both corroborates and conflicts with current theorizing on ostracism. Additionally, we provide practical insights into power estimation and effect moderation when designing studies using the Cyberball game. We declare no conflicts of interest, and potential reviewers that come to mind are Michael Bernstein, Ginette Blackhart and Jonathan Gerber, who have published earlier reviews on similar topics. Potential appropriate PLOS ONE academic editors are Harriet de Wit, Oscar García, and Michiel van Elk.

Please note we made all our research files available on the Open Science Framework (OSF). The link to this OSF page is provided in the manuscript. On this page we preregistered our hypotheses. Moreover, we presented parts of this meta-analysis at the EASP conference in Amsterdam last summer. Hence, it is likely reviewers will realize we are the authors. Personally, we do not consider this problematic.

We look forward to your reply, and hope you will find our research intriguing for review.

Kind regards,

Ilja van Beest,
also on behalf of Chris H.J. Hartgerink, Jelte M. Wicherts, Kipling D. Williams

Tilburg University
Warandelaan 2, 5037AB, Tilburg
i.vanbeest@tilburguniversity.edu

The Ordinal Effects of Ostracism:

A Meta-Analysis of 120 Cyberball Studies

Chris H.J. Hartgerink[1] ¶

Ilja Van Beest[2*] ¶

Jelte M. Wicherts[1]

Kipling D. Williams[3]

[1] Department of Methodology and Statistics, Tilburg University, North-Brabant, the Netherlands

[2] Department of Social Psychology, Tilburg Univeristy, North-Brabant, the Netherlands

[3] Department of Psychology, Purdue University, Illinois, United States of America

*Corresponding author

E-mail: i.vanbeest@tilburguniversity.edu

¶ These authors contributed equally to this work.

Abstract

We examined 120 Cyberball studies (N = 11,869) to determine the effect size of ostracism and conditions under which the effect may be reversed, eliminated, or small. Our analyses showed that (1) the average ostracism effect is large (d > |1.4|) and (2) generalizes across structural aspects (number of players, ostracism duration, number of tosses, type of needs scale), sampling aspects (gender, age, country), and types of dependent measure (interpersonal, intrapersonal, fundamental needs). Further, we test Williams's (2009) proposition that the immediate impact of ostracism is resistant to moderation, but that moderation is more likely to be observed in delayed measures. Our findings suggest that (3) both first and last measures are susceptible to moderation and (4) time passed since being ostracized does not predict effect sizes of the last measure. Thus, support for this proposition is tenuous and we suggest modifications to the temporal need-threat model of ostracism.

Keywords: Cyberball, meta-analysis, ordinal, ostracism

Introduction

Cyberball [1] is a virtual ball-tossing game that is used to manipulate the degree of social inclusion or ostracism in social psychological experiments. In this game the participant supposedly plays with two (or more) other participants, who are in fact part of the computer program. The program varies the degree to which the participant is passed the ball (see Fig. 1 for a still from the game). Ostracized players are not passed the ball after two initial tosses and thus obtain fewer ball tosses than the other players. Included players are repeatedly passed the ball and obtain an equal number of ball tosses as the other players. Our literature search showed that at least 200 published papers involved the use of the Cyberball paradigm to study ostracism and that over 19,500 participants have played the game thus far. In this paper we provide a meta-analysis of these studies. Our aim was to gauge the typical effect size of being ostracized in the Cyberball game and to see whether this effect is moderated by cross-cutting variables that were hypothesized to reduce/enhance the psychological impact of ostracism, structural aspects that are inherent in Cyberball (e.g., number of players, number of ball tosses), sampling aspects of the studies (e.g., gender composition), the type of dependent variables used (e.g., intrapersonal measures such as need satisfaction or interpersonal measures such as pro- or antisocial behavior), and the ordinal time point of the variable assessment (i.e., first or last).

Fig. 1. Cyberball game screenshot.

Historical background

Cyberball was introduced in 2000 as a means to study ostracism, that is: being excluded and ignored [1]. This focus of Cyberball on ostracism sets it apart from other paradigms that are tailored to study rejection, such as the future life rejection [2], the get-acquainted paradigm [3], and the autobiographical memory manipulation (i.e., remember a time when you were excluded [4]). The difference is that participants in Cyberball are not explicitly informed that they are excluded whereas in the other paradigms participants are provided a reason pertaining to why they are excluded. The Cyberball manipulation is a suitable method to study how people react to being ignored and excluded. Humans are social animals and care deeply about whether they are included or ostracized by others. Interestingly, ostracism is not only observed among loved ones, but on all levels of human organization. In fact, research suggests that most people are ignored and excluded at least once a day [3]. The social relevance is further evident in that ostracism not only affects the person who is ostracized (intrapersonal effects), but often also others (interpersonal effects). As a grim example, research on school shootings has suggested a direct link between ostracism and revenge. People who were ostracized may retaliate by murdering those responsible and sometimes even innocent bystanders [5]. The impact of ostracism is also evident in research findings using Cyberball. Through experimental work, it has been repeatedly shown that being ostracized has an effect on people—either on their psychological functioning (e.g., decreases in positive mood [6]) or on certain interpersonal behaviors (e.g., increases in social susceptibility or aggressive behaviors [7,8]). These experiments have highlighted the (mostly negative) impact of ostracism on fundamental needs (e.g., belonging [9]), mood, physiology (e.g., body temperature [10]), and various other constructs, including those measured with behavioral measures (e.g., conformity, compliance, aggression). In the current paper, we refer to the general effect of being ostracized compared to being included in Cyberball as the *ostracism effect*.

To capture how people respond to ostracism, Williams [11] proposed a temporal need-threat model of ostracism. Here he suggested three stages of the ostracism effect, namely: (1)

a *reflexive* stage, (2) a *reflective* stage, and (3) a *resignation* stage. In the reflexive stage, the response to the ostracism sequence is immediate and occurs like a reflex. This initial response is theorized to be socially painful, threatening [9] and, following overdetection theory [12], should be easily detectable due to evolutionary over-sensitivity to cues of ostracism. Such a reflex would not take into account situational specifics and provides little room for coping. The reflex is proposed to affect primarily pain, fundamental needs, and emotional reactions (e.g., increased anger and sadness). The affected fundamental needs are belonging, self-esteem, control, and meaningful existence, typically measured by a need satisfaction scale [11]. According to Williams, measures of reflexive responses must occur during, or in the case of self-report measures, immediately following Cyberball (with the wording of the questions referring to how participants felt *during the game*). The *reflective* (or delayed) stage, which follows this immediate response, is subject to more rational thought and coping with the threats. Part of such coping is the necessity for fortification of the threatened fundamental needs. Coping can be measured both in terms of speed of recovery (higher levels of need satisfaction approaching the levels of included participants) and emotional, cognitive, and behavioral choices. The *resignation* stage occurs after prolonged ostracism, causing prolonged periods of pain and more fundamental need threat. If one is not able to fortify the fundamental needs, a prolonged ostracism sequence leads to feelings of helplessness, alienation, depression, and unworthiness. Because the resignation stage is hypothesized to occur only after prolonged and repeated exposure to ostracism (as in months or years), it is not feasible (and even unethical) to study resignation responses in laboratory experiments. Hence, in this paper we limit ourselves to studying the reflexive and reflective stages. For these stages, Williams asserts that moderation and variation of need satisfaction effects by individual differences and socially relevant factors (e.g., type of group from which one is excluded) will be less likely to occur for reflexive measures than for reflective measures.

Goals of meta-analysis

A limited number of Cyberball experiments have been reviewed in other meta-analyses, but these meta-analyses had a different goal than the current meta-analysis. Previous meta-analyses focused on social rejection and not on ostracism [13,14], or focused only on a specific dependent variable (e.g., fMRI [15,16]). Importantly, none of these early meta-analyses were specifically set up to test Cyberball effects only. Consequently, we do not know how structural variables of Cyberball or sample characteristics affect the ostracism effect size. Moreover, none of these meta-analyses considered whether it matters if a specific variable is measured first or last. Thus, it remains unclear whether the ostracism effect size decreases or increases over time and whether immediate measures are more or less moderated by cross-cutting variables. The goal of our meta-analysis is to provide a comprehensive understanding of the Cyberball-induced inclusion versus ostracism effect size. Under what conditions, if any, is the effect size negative, zero, or especially small? Under what conditions is it especially large? To answer these questions we made several selection decisions (see also the Open Science Framework (OSF) where we preregistered all selections and hypotheses).[1]

The first selection decision is that we considered only the first and the last dependent variable of all included studies. The reason for this selection was that it allowed us to gauge whether the effect sizes are affected by the time point at which the effects are measured. Another reason is that it served as a proxy to evaluate the hypothesis that immediate measures should be less affected by cross-cutting variables than more delayed measures.

A second decision is that we considered two different approaches to test whether first and last measures can be moderated by cross-cutting variables. This allowed us to test the robustness of our hypothesis across independent variables. The first approach to assess moderation was to conduct a meta-analysis on all studies that were explicitly designed to test whether being ostracized or included can be moderated by a cross-cutting factor. For this purpose we selected all the studies that included an experimentally manipulated moderator variable. Moreover, to meta-analyze the interaction term for first and last measure we followed the prediction of the authors in computing this interaction term. A potential limitation of our decision to follow the prediction of the authors is that the predictions may have been generated post-hoc on the basis of observed outcomes. For example, if authors used a 2 (ostracized vs included) x 2 (ingroup vs outgroup design) we followed the prediction of the authors to compute whether the interaction term denotes that ostracism is increased by an outgroup or decreased by an outgroup (specific calculations are reported in the methods section and formulae in the Appendix). Moreover, after computing the overall interaction terms we created dotplots in which we depicted the effect of ostracism across the two levels of the moderator and – perhaps more importantly - the effect of the moderator across the two levels of the ostracism manipulation. This was done to facilitate the interpretation of an interaction term and specifically to show whether cross-cutting variables have more impact on being included in Cyberball or more impact on being ostracized in Cyberball [17].

The second approach to test moderation was to assess if and how first and last measures are moderated by structural aspects of Cyberball (i.e., number of depicted Cyberball players, number of ball tosses used, duration of the game) and sample aspects (i.e., gender composition, country of origin, age). Note that the outcome of this analysis may thus also be used for future researchers to decide how to set up a game of Cyberball and whether effects generalize across age, gender, and country of origin. Because prior research has not explicitly manipulated structural aspects in controlled experiments we did not have a specific prediction whether increasing the number of players, ball tosses, and game duration would increase or diffuse the impact of ostracism. Given that the social aspects of an interdependent setting may be less evolutionary relevant for males than for females [18] and less relevant for older people than younger people [19], we explored whether an increase of male participants and mean age would decrease the ostracism effect. Moreover, considering that collectivism might influence the degree to which belonging is important [20], we used a categorization of continents (i.e., U.S., other western countries, Asian countries, and remaining countries) to explore whether a more collective orientation would be associated with larger ostracism effects. Finally, because some of the factors might be related (i.e., an increased number of ball tosses is likely to be associated with an increase in duration), we decided to use a regression approach in which all factors were entered simultaneously. A benefit of this approach is that it ensures that significant predictors have an impact above and beyond the impact of the other predictors.

The third decision is that we also checked the robustness of our findings across various dependent variables. More specifically, we coded whether the first and last measures belonged to the category of *interpersonal* variables assessing how ostracism impacts others or belonged to the category of *intrapersonal* variables assessing how ostracism impacts the self. Examples of interpersonal measures are donations to charity, helping behavior, money allocations in economic games, and aggression measures such as irritating sounds blasts or hot sauce allocation. These were initially coded into pro- and anti-social, but were collated into the category interpersonal due to small k the first measure (4 and 10, respectively) and last measure (8 and 6, respectively). Examples of intrapersonal measures are self-reported anger, self-esteem, control, and physiological measures such as body temperature or galvanic skin response. A benefit of classifying all variables into broad categories is that it increases the power of the meta-analysis since expanding the analysis to even more specific constructs would seriously limit the number of available studies. We made one exception and that is that we also ran tailored analyses on a subset of the intrapersonal measures that assessed *fundamental needs* (i.e., belonging, self-esteem, control, and meaningful existence). These fundamental needs measures included the typical need satisfaction measures that are especially designed for Cyberball [1,21,22] and conceptually related measures such as the

Rosenberg Self-Esteem Scale. The reason why we did focus on this specific subset of intrapersonal variables is that the evidence supporting Williams' temporal model is to a large extent based on studies using these specific dependent variables. In other words, these fundamental needs measures are particularly important for testing Williams's [11] prediction concerning moderation of ostracism effects over time.

Hypotheses

Following our preregistered report on OSF, we divided the hypotheses into two primary hypotheses and several secondary hypotheses. The two primary hypotheses were: is there an ordinal decrease of the ostracism effect across time of measurement? (Hypothesis 1) and is there an ordinal difference in the interaction effect across time of measurement (Hypothesis 2)? Secondary hypotheses regarded moderation of the ostracism effect by structural aspects of the studies, sampling aspects of the studies, and different types of dependent measures used. These hypotheses will be answered with random and mixed-effects meta-analytic models applied to all 120 studies that we were able to collate.

Method

Study inclusion criteria

First, we only considered Cyberball experiments that contained a factor that manipulated the number of virtual ball tosses obtained by the participants. For this ostracism factor we only considered the condition in which participants were ostracized by all other participants and the condition in which participants were equally included by all other players. Second, we only considered experiments that incorporated a between-subjects design with random assignment. Within-subject designs were excluded, because this would require the correlations between measures in primary studies and such correlations are often not reliably reported in the papers. Moreover, most within-subjects designs regard high-dimensional neurophysiological measurements such as fMRI that are beyond the scope of this meta-analysis [15,16]. Third, we checked whether the experiments contained other factors besides the ostracism factor. If the experiment contained more than two additional factors we collapsed effects sizes across the factor that authors expressed least interest in. Moreover, continuous variables that were dichotomized into factorial levels were also collapsed due to the many problems dichotomization can cause (e.g., underestimation of effect size, spurious effects [23,24]; four cases). Fourth, for the dependent measures the criterion was that they were (expected to be) affected by the ostracism manipulation. We considered the measures that immediately followed the manipulation (first measure) and the measure at the end of the study (last measure), while excluding manipulation checks in this assessment.

Reasons for these inclusion criteria are threefold: (1) Most Cyberball experiments take place in such a format, making it an encompassing criterion for the purposes of this meta-analysis. (2) The choice to limit the meta-analysis to between-subject designs rendered computational aspects more feasible based on reported statistics in papers. (3) The criteria maximize experimental rigor as they minimize the need for subjective quality assessment of the primary studies. Indeed, clear inclusion criteria decrease variability due to design characteristics, which increases power for moderator analyses [25].

Literature search

To have a comprehensive meta-analysis of Cyberball studies, we used seven search strategies in the period of November 2012 through April 2013. These search strategies included database searches, a call for data, cross-reference with Kip Williams's online list of Cyberball studies, Google Scholar alerts, citation records, Society for Personality and Social Psychology (SPSP) conference abstracts, and personal communications.

The databases searched included Web of Knowledge, PubMed, ScienceDirect, and Worldcat using all sources from the Tilburg University library. The first three cover only published articles, whereas Worldcat also covers books and dissertations as well as the PsycINFO database. All these databases were searched with the keywords *cyberball, ball-tossing* and *ball AND ostraci**. Web of Knowledge was the first database searched. For this database, an additional search term (i.e., *ball AND exclu**) was used, but this additional search term yielded zero relevant hits that were not a result of the other searches and was dropped. Across all these searches, results included 1927 potentially relevant studies of which a total of 109 were deemed relevant and saved for coding. Within Web of Knowledge, we looked through all citation records of the seminal papers by Williams et al. [1]; Williams and Jarvis [26]. These papers were cited 332 times (as of 5th of November, 2012), of which 43 papers were saved for coding. The entire literature search provided 2259 potentially relevant studies (including possible duplicates across searches), of which 152 were selected to be included in the coding.

The call for data was put on the list servers or forums of SPSP, European Association of Social Psychology (EASP), and Social Psychology Network (SPN; all on 3rd of December, 2012). This resulted in 9 replies, yielding 3 useful studies.

Kip Williams keeps a list of Cyberball studies on his website. This list was used to check for extra articles that did not turn up in the initial searches on November 15th, 2012.[2] The list included 93 papers, of which 9 papers were included to be coded.

The final searches included Google Scholar alerts, SPSP conference abstracts, and personal communication. The Google Scholar alerts were used to keep up to date with new literature. These alerts notify a user when new search results for a search term occur and were used for *cyberball* and *ball-tossing.* This yielded 85 search results of which 25 were saved for coding. SPSP conference abstracts from 2006 through 2013 were searched for Cyberball studies. This led to personal communications with the authors of the conference abstracts, leading to additional studies. Pooled, the personal communication and the conference abstracts yielded 21 potentially relevant studies, of which 20 were saved for coding. The seminal paper by Williams et al. [1] was added separately.

In sum, the literature search spanned 2468 potentially relevant studies, resulting in 205 that were saved for coding. During coding, papers were assessed to fit the inclusion criteria. Of the 205 papers, 107 papers were excluded for a variety of reasons. See also Fig. 2. Several involved the use of a within-subjects design (52 papers). Some papers could not be accessed (5 papers) or could not be included because we did not receive the required data on request (7 papers). Some were excluded for other reasons (43 papers), such as not involving new data (e.g., a dissertation study that was later published). All included papers were published between 2000 (after the introduction of Cyberball) and April 2013. This resulted in a final, fully coded sample of 98 papers containing 120 studies, with mean sample size 98.9 and median sample size 74.[3] There were a total of 11,869 Cyberball participants.

Fig. 2. PRISMA flowchart of the current meta-analysis.

Coding procedure

The first author coded all the studies and conducted all the analyses. The second author double-checked the coding of all 52 studies that entailed a full two-by-two design. The third author double-checked and reran the R code of all analyses. Finally, an extensive account of all coding decisions is publicly available via Open Science Framework on a paper-by-paper basis (see Footnote 2 for the direct link, Supplement S1 also contains the data).

We first coded the structural aspects and sample aspects of all papers. The structural aspects of Cyberball that we coded were (1) number of players depicted in Cyberball, (2) total number of ball tosses used throughout the game, (3) total duration of the game in seconds.

The sample aspects that we coded were (1) percentage of male participants, (2) average age of participants, and (3) country of origin.

We then coded the dependent variables that were relevant for the current meta-analysis by retrieving the means and standard deviations of the first and the last relevant measure of all papers. Importantly, to estimate the duration between the first and last measure we counted the number of questions that were assessed between the two measures. Specifically, following a longstanding practice in the freshman testing program of the University of Amsterdam [27] we estimated that participants would need 6 seconds on average to complete one question. Moreover, we included additional time if this was explicitly reported in the method section of the manuscript or when a measure would clearly deviate from 6 seconds to complete (e.g., tasks that measure endurance such as a grip strength task).

Both first and last measures were subsequently coded in the following general terms: (1) interpersonal, (2) intrapersonal, (3) fundamental needs, (4) model correspondence. Interpersonal measures were defined as measuring constructs that relate to (the self and) others (e.g., *how angry do you feel towards person X?*, donations to charity). Intrapersonal measures were defined as measuring constructs that relate only to the self (e.g., *how angry do you feel?*, physiological measures). Fundamental needs measures were those that measured self-esteem, belonging, control, meaningful existence, or a composite of these. Note that the fundamental needs are a refinement of the intrapersonal measures and that intrapersonal measures thus include the fundamental need measures. The model correspondence variable coded whether the first- and last measure fit the definition William's ostracism model that a variable can indeed be classified as an immediate measure (i.e., during the game) and delayed measure (i.e., after the game/now), respectively.

The consequence of including many different kinds of dependent variables is that some measures are expected to increase as a function of ostracism (e.g., need threat) and others are expected to decrease (e.g., need satisfaction). To counteract computational problems (i.e., cancellation of effects) being caused by this bidirectionality of ostracism effects, we coded the direction of the ostracism effect for each specific measure, such that negative effect sizes depict negative psychological effects.

A similar argument can also be made about including multiple moderator variables in the analysis of interaction effects. In the 52 studies that included a moderator variable we thus needed to account for the expected direction of every moderator. If we had not done this, the interaction effects could cancel out, thereby leading to ambivalent results. To explain this, we present in Table 1 hypothetical data for the four different study designs that are possible when crossing direction of the effect and direction of the moderation. The relevant effect sizes should be corrected to attain comparable effect sizes across studies. Effect sizes for the simple ostracism effect (column wise) were corrected only for the type of measure. For instance, for panels (a) (involving, e.g., need threat) and (c) (involving, e.g., need satisfaction), the corrections entailed a multiplication with -1 or +1, respectively. Simple moderator effects (row wise comparisons) are interesting for understanding the effect of the moderator under either ostracism or inclusion. These simple moderator effects were corrected for both the type of measure *and* the expected moderation (i.e., exacerbation, -1, or minimization, +1). For example in panel (c), the 5 and 8 on the right are used to compute the *standard ostracism effect* (as in [1])*,* whereas the 3 and 8 in the left column represent an ostracism effect that is thought to be exacerbated. For example, in a given ostracism study with a two-by-two design, adolescents are expected to show stronger ostracism effects, compared to young adults [19]. The 5 and 8 would subsequently represent the scores for the young adults, whereas the 3 and 8 would represent the scores for the young adolescents. In panel (d) we depict a study in which the *moderated* column is thought to lead to a minimal ostracism effect, as could be expected when Cyberball is played with members of a despised out-group [28]. The margins (greyed out) denote the simple effects, which are after correction comparable across all panels (a) through (d), indicating that this correction did what we intended it to.

Table 1. Hypothetical data example of coding correction.

(a) Negative moderator, negative measure

		Moderated	Not-moderated/control	Raw	Correct
Ostracism factor	Ostracism	13	11	2	2
	Inclusion	8	8	0	0
	Raw	5	3		
	Correct	-5	-3		

(b) Positive moderator, negative measure

		Moderated	Not-moderated/control	Raw	Correct
Ostracism factor	Ostracism	9	11	-2	2
	Inclusion	8	8	0	0
	Raw	1	3		
	Correct	-1	-3		

(c) Negative moderator, positive measure

		Moderated	Not-moderated/control	Raw	Correct
Ostracism factor	Ostracism	3	5	-2	2
	Inclusion	8	8	0	0
	Raw	-5	-3		
	Correct	-5	-3		

(d) Positive moderator, positive measure

		Moderated	Not-moderated/control	Raw	Correct
Ostracism factor	Ostracism	7	5	2	2
	Inclusion	8	8	0	0
	Raw	-1	-3		
	Correct	-1	-3		

Raw denotes the simple effect in the hypothetical data before correction whereas correct denotes the simple effect after correction. Column wise effects are multiplied by the type of measure only, whereas row wise effects are multiplied by both the type of moderator and type of measure.

Finally, relevant information that was missing in the papers was requested from the authors via e-mail. In case of non-response, we sent three follow-up e-mails. All this communication was documented and can be found on the OSF page for this project. In case of non-response or non-willingness to send data, studies were either eliminated if the information was crucial (i.e., means and standard deviations of the measures per group), computed if possible (i.e., cell sizes), or assumed if deemed reasonable on the basis of additional information. For instance, when no information was given we considered the Cyberball manipulation characteristics to be similar to previous studies in the same paper or in earlier papers referred to in the paper (descriptions of all cases are described in the log file on the OSF).

Statistical analyses

For the analyses, we used version 1.9-5 of the *metafor* package [29] in the R statistical environment [30].

Effect size metric

We used Hedges's *g* version of the standardized mean differences as the effect size. Hedges's *g* corrects for the slightly biased estimate given by Cohen's *d* [31]. Standardized effects were calculated across the ostracism factor, where the 52 studies with a cross-cutting variable were included as a simple effect of ostracism within the non-moderated level. Standardized interaction effect were calculated by taking the standardized difference between the unstandardized main effects (see the Appendix for the exact formulae used). These effects were computed for both the first and last dependent variable in each experiment. For example, in a 2 (ostracized vs. included) by 2 (moderator present vs. moderator absent) design with multiple measures, we calculated two simple ostracism effects (Hypothesis 1) and two interaction effects (Hypothesis 2). For ten studies, more factors/levels were used and a 2 by 2 was extracted.

Meta-analytic model

We used random- and mixed-effects models, because heterogeneity in the effect sizes is expected due to both the inclusion of different measures and additional unknown methodological and substantive factors. The meta-regression element in some of the analyses is the variable time as predictor of the ostracism effect. Analyses without this study-level predictor reduce to a random-effects model. We used Restricted Maximum Likelihood (REML) to estimate tau-squared (i.e., the residual variance), as recommended by Viechtbauer [32]. Note that when estimating a mixed- or random effects model, one does not estimate a single *true* effect, but rather the mean and variance of underlying effects [32].

Statistical sensitivity analyses

To test for robustness of the effects, we incorporated several statistical sensitivity analyses. We flagged possibly problematic outliers on the basis of studentized deleted residuals, Q-Q plots, and Cook's distance values. Subsequently, we inspected the effect of these outliers on substantial results in statistical sensitivity analyses in which these outliers were excluded. Another statistical sensitivity analysis entailed fitting of the mixed-effects model with tau-squared fit at the upper bound value of the 95% confidence interval.

Funnel plot asymmetry

A funnel plot depicts each study's effect size against its standard error [33]. Larger studies have smaller standard errors, and vice versa for smaller studies. Following from a theoretical fluctuation of the population effect size due to sampling variance, a funnel plot should be symmetrical around the estimated mean effect size. If there are no methodological or substantive reasons to expect a link between effect sizes and standard errors, funnel plot *asymmetry* can indicate publication bias (e.g., [34]). To test funnel plot asymmetry, we used Egger's regression test [35] for mixed-effects models [36].[4] This tests whether the distribution of effect sizes is equal on both sides of the average effect, when accounting for true heterogeneity. Funnel plot asymmetry thus indicates bias in the estimated mean effect size and possibly publication bias.

Results

In our reporting of the effect sizes, *d* indicates a main effect and Δ*d* indicates an interaction effect. Even though we used Hedges's *g*, we maintained the notation of *d*, because *g* is only a minor correction to Cohen's *d*. Statistical sensitivity analyses are only reported if they showed different effects (all statistical sensitivity analyses can be found on OSF).

Primary analyses

The two primary hypotheses are tested in four meta-analyses, of which the study level effects are reported in Table 2. The table includes effect sizes used in the estimation of the average simple effect of ostracism on the first measure, the average simple effect on the last measure and the estimation of the average interaction effect on both the first and last measure.

Table 2. Effect sizes per study for the primary hypotheses.

First author	Year	*N*	*d* T1	(*SE*)	*d* T2	(*SE*)	Δ*d* T1	(*SE*)	Δ*d* T2	(*SE*)
Alvares	2010	74	-1.21	0.12	-0.10	0.10	-0.15	0.24	1.12	0.23
Ambrosini	2013	40	-1.69	0.13	-0.97	0.11	-	-	-	-
Aydin	2012	68	-0.95	0.13	-0.40	0.12	-1.19	0.24	0.72	0.23
Banki	2012	89	-1.87	0.07	-0.35	0.05	-	-	-	-
Bastian	2010	72	-2.75	0.11	-1.42	0.07	-	-	-	-
Bernstein	2012	24	-0.41	0.16	-	-	-	-	-	-
Bernstein	2012	25.50	-1.04	0.17	-	-	-	-	-	-
Bernstein	2010	73	-1.63	0.16	-1.63	0.16	-0.86	0.37	-1.11	0.40
Bernstein	2010	138	-2.67	0.10	-1.96	0.08	-0.53	0.22	-0.51	0.17
Bernstein	2012	67	-2.00	0.17	-0.99	0.13	-1.07	0.45	-0.80	0.30
Bernstein	2012	27	-1.39	0.17	-	-	-	-	-	-

First author	Year	*N*	*d* T1	(*SE*)	*d* T2	(*SE*)	Δ*d* T1	(*SE*)	Δ*d* T2	(*SE*)
Boyes	2009	89	-0.43	0.05	-0.80	0.05	-	-	-	-
Boyes	2009	87	-0.20	0.05	-0.84	0.05	-	-	-	-
Brochu	-	35	-2.51	0.20	-0.48	0.11	-	-	-	-
Brown	2009	52	-0.64	0.08	-	-	-	-	-	-
Carter	2008	143	-0.28	0.06	0.20	0.06	0.34	0.11	0.17	0.11
Carter-Sowell	2008	65	-2.86	0.12	-1.48	0.08	-	-	-	-
Carter-Sowell	2010	74	-1.60	0.14	-1.49	0.13	-1.23	0.33	-1.15	0.34
Carter-Sowell	2010	70.67	-2.09	0.17	-0.56	0.11	-0.65	0.39	-0.63	0.24
Chen	2012	60	-1.04	0.14	-	-	-1.35	0.27	-	-
Chen	2012	83	-1.32	0.11	-	-	-1.32	0.21	-	-
Chernyak	2010	76	-1.52	0.10	0.15	0.08	-	-	-	-
Chow	2008	75	-1.20	0.06	-1.31	0.06	-	-	-	-
Chrisp	2012	77	-0.70	0.06	-0.15	0.05	-	-	-	-
Coyne	2011	40	-0.56	0.10	-	-	-	-	-	-
De Waal-Andrews	2012	136	-3.55	0.16	-2.55	0.11	-1.29	0.24	-0.87	0.18
De Waal-Andrews	2012	112	-4.21	0.22	-2.17	0.11	-1.56	0.31	-1.20	0.18
DeBono	-	57	-1.07	0.15	-0.05	0.13	-1.55	0.29	-0.48	0.27
DeBono	-	81	-1.07	0.11	-0.10	0.09	-0.33	0.21	0.24	0.19
DeBono	-	83	-0.13	0.09	-	-	-0.75	0.19	-	-
Dietrich	2010	75	1.43	0.07	-	-	-	-	-	-
Duclos	2012	59	-0.63	0.07	-	-	-	-	-	-
Eisenberger	2006	48	-0.15	0.08	-1.24	0.10	-	-	-	-
Fayant	-	60	-2.04	0.20	-1.12	0.15	0.22	0.38	-0.44	0.28
Floor	2007	88	-1.92	0.13	-0.73	0.09	-0.21	0.28	-0.59	0.19
Gallardo-Pujol	2012	57	-1.18	0.16	-0.52	0.15	-1.17	0.31	0.11	0.29
Gan	2012	72	-0.54	0.03	-0.07	0.03	-0.62	0.06	0.02	0.06
Garczynski	2013	83	-1.51	0.19	0.39	0.15	-1.29	0.33	-0.01	0.29
Geniole	2011	74	0.19	0.06	-0.11	0.06	-	-	-	-
Gerber	-	38	-2.09	0.16	-	-	-	-	-	-

First author	Year	*N*	*d* T1	(*SE*)	*d* T2	(*SE*)	Δ*d* T1	(*SE*)	Δ*d* T2	(*SE*)
Gerber	-	89	-3.38	0.21	-	-	-	-	-	-
Gonsalkorale	2007	97	-1.31	0.14	0.26	0.12	0.49	0.30	1.31	0.25
Goodwin	2010	300	-1.81	0.04	-0.94	0.03	0.20	0.08	-0.43	0.07
Goodwin	2010	314	0.13	0.02	-0.09	0.02	0.35	0.06	-0.10	0.06
Greitemeyer	2012	56	-0.48	0.07	-0.23	0.07	-	-	-	-
Gruijters	-	113	-0.26	0.06	-1.07	0.07	-	-	-	-
Hackenbracht	2013	51	-1.92	0.11	-0.18	0.08	-	-	-	-
Hawes	2012	55	-2.16	0.23	0.69	0.15	0.00	0.38	-1.05	0.28
Hellmann	-	76	-1.21	0.12	0.19	0.10	-1.40	0.22	0.74	0.21
Hess	2010	162	-2.34	0.04	-0.87	0.03	-	-	-	-
Hess	2011	38	-0.64	0.11	-	-	-	-	-	-
Horn	-	68	-0.77	0.12	-0.99	0.13	-0.99	0.23	1.49	0.24
IJzerman	2012	86	-1.67	0.12	-	-	-1.07	0.22	-	-
Jamieson	2010	33	-1.56	0.15	-1.06	0.13	-	-	-	-
Jamieson	2010	68	-1.94	0.09	-1.47	0.07	-	-	-	-
Johnson	2010	104	-0.73	0.04	-0.79	0.04	-	-	-	-
Kassner	-	85	-1.72	0.13	-1.02	0.11	-0.87	0.31	-0.30	0.21
Kassner	2012	49	-2.11	0.12	-1.78	0.11	-	-	-	-
Kerr	2008	250	-1.66	0.02	-0.05	0.02	-	-	-	-
Kesting	2013	76	-0.28	0.05	-0.79	0.06	-	-	-	-
Knowles	2010	62	-0.38	0.12	-	-	-0.99	0.25	-	-
Knowles	2012	60	-0.60	0.07	-	-	-	-	-	-
Krijnen	2008	144	-4.74	0.11	-0.18	0.03	-	-	-	-
Krill	2008	119	-2.11	0.05	-0.57	0.03	-	-	-	-
Lakin	2008	36	-1.53	0.14	-0.51	0.11	-	-	-	-
Lau	2009	56	-2.50	0.23	-1.09	0.15	-0.06	0.58	1.36	0.46
Lustenberger	2010	71	-0.83	0.06	0.04	0.06	-	-	-	-
Lustenberger	2010	156	-0.70	0.03	-	-	-	-	-	-
MacDonald	2008	63	-0.15	0.06	-	-	-	-	-	-

First author	Year	*N*	*d* T1	(*SE*)	*d* T2	(*SE*)	Δ*d* T1	(*SE*)	Δ*d* T2	(*SE*)
McDonald	2012	270	-0.06	0.02	-2.40	0.03	-	-	-	-
Nordgren	2011	71	-0.74	0.06	-	-	-	-	-	-
Nordgren	2011	74	-0.80	0.06	-	-	-	-	-	-
Nordgren	2011	46	-2.24	0.14	-	-	-	-	-	-
Nordgren	2011	44.67	-0.55	0.09	-0.75	0.09	-	-	-	-
Nordgren	2011	58.67	-0.65	0.07	-	-	-	-	-	-
Oberleitner	2012	88	-2.36	0.08	0.42	0.05	-	-	-	-
O’Brien	2012	125	-0.58	0.03	-0.69	0.03	-	-	-	-
Peterson	2011	40	-0.89	0.11	-0.91	0.11	-	-	-	-
Pharo	2011	74	-1.33	0.13	-0.58	0.11	-1.01	0.30	-0.84	0.23
Plaisier	2012	149	-0.36	0.05	0.23	0.05	-0.40	0.11	-0.56	0.11
Ramirez	2009	121	-2.26	0.05	-1.02	0.04	-	-	-	-
Ren	2012	53	-2.18	0.12	-0.17	0.07	-	-	-	-
Renneberg	2011	60	-1.46	0.16	-1.30	0.15	0.47	0.29	0.51	0.29
Riva	2011	100	-2.10	0.13	-1.09	0.09	-	-	-	-
Ruggieri	-	91	-0.39	0.04	-0.57	0.05	-	-	-	-
Ruggieri	-	74	-0.06	0.13	-0.23	0.13	-0.31	0.24	-0.68	0.23
Sacco	2011	51	-2.40	0.13	-1.45	0.10	-	-	-	-
Sacco	2011	21	-2.28	0.29	-1.46	0.22	-	-	-	-
Sacco	2011	38	-1.74	0.14	-1.04	0.11	-	-	-	-
Salvy	2010	59	-1.45	0.08	-1.43	0.08	-	-	-	-
Salvy	2009	103	-1.48	0.05	-1.31	0.05	-	-	-	-
Schaafsma	2012	720	-1.42	0.02	-0.49	0.02	0.09	0.03	0.33	0.03
Segovia	2012	56	0.14	0.13	-	-	-1.89	0.32	-	-
Staebler	2011	68	-0.79	0.12	-0.05	0.12	0.50	0.23	0.42	0.23
Stillman	2009	121	-0.74	0.15	-1.13	0.16	0.57	0.22	-1.19	0.24
Stock	2011	155	-2.00	0.04	-0.13	0.03	-	-	-	-
Van Beest	2011	87	-0.94	0.10	-0.58	0.09	-0.40	0.24	-0.44	0.19
Van Beest	2011	183	-2.64	0.13	-0.50	0.07	-0.76	0.22	-0.11	0.13

First author	Year	*N*	*d* T1	(*SE*)	*d* T2	(*SE*)	Δ*d* T1	(*SE*)	Δ*d* T2	(*SE*)
Van Beest	2006	135	-1.29	0.07	-0.65	0.06	-0.10	0.14	-0.13	0.12
Van Beest	2006	111.33	-2.11	0.11	0.09	0.07	-0.09	0.22	-0.19	0.14
Van Beest	2012	125	-2.68	0.11	-1.24	0.07	0.06	0.35	-0.23	0.15
Van Beest	2012	85	-3.10	0.20	0.05	0.09	-0.28	0.44	0.07	0.18
Van Beest	2013	49	-3.97	0.24	-1.32	0.10	-	-	-	-
Van Beest	2013	91	-3.17	0.20	-0.48	0.09	0.75	0.56	0.53	0.18
Van Dijk	-	51	-1.50	0.10	-0.04	0.08	-	-	-	-
Webb	-	170	-0.91	0.05	-0.38	0.05	0.03	0.10	0.04	0.09
Weik	2010	65	0.16	0.12	-0.22	0.12	-0.43	0.24	0.66	0.24
Wesselmann	2009	82	-0.71	0.10	-2.03	0.14	-1.30	0.24	-0.20	0.28
Wesselmann	2012	91	-1.46	0.06	-	-	-	-	-	-
Williams	2002	390	-0.39	0.01	-2.35	0.02	-	-	-	-
Williams	2000	732	-0.79	0.01	-1.44	0.01	-	-	-	-
Williams	2000	111	-0.26	0.06	-1.01	0.07	-0.20	0.15	-0.98	0.15
Wirth	2009	159.33	-2.29	0.08	-0.76	0.05	0.05	0.17	0.46	0.11
Wirth	2010	76	-0.96	0.06	-1.64	0.07	-	-	-	-
Zadro	2004	62	-1.63	0.16	-0.19	0.12	-0.11	0.32	-1.12	0.28
Zadro	2004	77	-1.75	0.14	-0.33	0.10	-0.29	0.28	-0.70	0.21
Zadro	2006	56	-3.70	0.19	-0.87	0.08	-	-	-	-
Zhong	2008	52	-0.72	0.15	-	-	-	-	-	-
Zoller	2010	57	-0.24	0.07	-0.09	0.07	-	-	-	-
Zwolinski	2012	56	-2.01	0.11	-0.28	0.07	-	-	-	-

d T1 refers to ostracism effect on first measure; *d* T2 refers to ostracism effect on last measure; Δ*d* represent interactions. Multiple rows for the same first author and year is possible due to multiple studies across papers. Non-integer *N*s arise from division of full sample *N* for included conditions, appropriate due to random assignment (e.g., two conditions out of 3, when sample is 56: (56 / 3) × 2 = 37.333). Supplement S2 gives the full reference list of the papers in this table.

Simple ostracism effect (Hypothesis 1)

In a random-effects model on the main effect of ostracism (k = 120), residual heterogeneity was significant, Q (119) = 1395, p < .001, I^2 = 92.99% and estimated at τ^2 = 0.90, 95% CI [0.70, 1.24]. The heterogeneity measure τ^2 includes both the estimated proportion of explained variance at the study level and unexplained variance in the distribution of underlying effect sizes (i.e., τ_{res}^2). The analysis yielded an estimated average effect of d = -1.36, p < .001, 95% CI [-1.54, -1.18]. A random-effects version of the Egger's test [36] indicated funnel plot asymmetry, Z = -6.14, p < .001. Due to the size of the average effect, hence large power to acquire significant outcomes in primary studies, we do not suspect publication bias to explain this asymmetry. In other words, immediately after being ostracized, the average ostracism effect is estimated at -1.36 standard deviation units, which entails a large effect [37].

Next, we fitted a mixed-effects regression model for the ostracism effect on the last measure (k = 95), including estimated time in seconds since completing the Cyberball game as predictor. Residual heterogeneity was significant, Q_E (93) = 803, p < .001 and estimated at τ_{res}^2 = 0.38, 95% CI [0.27, 0.54]. The intercept was estimated at $d_{intercept}$ = -0.76, p < .001, 95% CI [-0.91, -0.61]. Moreover, the estimated time in seconds between exclusion in Cyberball and the moment at which the last measure was taken failed to moderate the average effect, b = 0.0069, p = .187, 95% CI [-0.0034, 0.0172]. However, we have to take into consideration the low power of the moderation analyses due to the large (residual) heterogeneity in effect sizes [25]. A regression test for mixed-effects model with moderator (i.e., including both the time and SE as predictor) showed no funnel plot asymmetry, Z = -0.72, p = .474. In short, long after ostracism has occurred (M_{time} = 4.85 minutes), ostracized participants on average scored around -0.73 standard deviation units lower when compared with included participants, an effect that does not appear to be moderated further by time passed since the ostracism occurrence.

Thus, results show a clear effect of ostracism on both the first and last measures, of which the latter is *not* predicted by our operationalization of time. The ostracism effect over time can also be inspected via confidence intervals. Comparing the 95% confidence intervals for the average ostracism effect on the first measure (i.e., [-1.54, -1.18]) and on the last measure (i.e., [-0.86, -0.59]) showed no overlap. Although the difference in average effect sizes between first and last measure cannot be formally tested (because of a lack of information on the correlation between measures in the primary studies), the mean difference is sizeable and CIs confirms our prediction that the average ostracism effect is smaller for the last measure. In fact, given the expected positive correlation between effects for first and last measures, the comparison of CIs is likely to be conservative [38]. Additionally, we noted that estimated residual heterogeneity was larger on the first- than on the last measure. We conclude that the average ostracism effects decreases from the first- to last measures and that study-level effects are more similar on the last measure.

Moderation of ostracism (Hypothesis 2)

To test moderation of the ostracism effect, we selected the factorial experiments that manipulated ostracism and another independent variable in between-subjects designs. A random-effects model on the interaction effect (Δd) on the first measure (k = 52) showed heterogeneity in underlying effects, Q (51) = 103.24, p < .001, I^2 = 50.60% and an estimated τ^2 = 0.19, 95% CI [0.07, 0.41]. The average interaction effect equaled Δd = -0.46, p < .001, 95% CI [-0.64, -0.28], indicating a change in the ostracism effect due to the moderator level and vice versa (i.e., moderation of the ostracism effect). There was indication of funnel plot asymmetry in this analysis, Z = -2.43, p = .015. Thus, the data indicate that, across the board, the ostracism effect *can* be moderated on the first measure following the ostracism sequence, but it is possible that publication bias may have affected the interaction estimates.

On the last measure (k = 46), the mixed-effects model (with estimated time as predictor) for the interaction effect again showed residual heterogeneity, Q_E (44) = 100.82, p < .001 and estimated τ_{res}^2 = 0.21, 95% CI [0.10, 0.55]. The intercept of the interaction effect was estimated at $\Delta d_{intercept}$ = -0.20, p = .052, 95% CI [-0.402, 0.002] and no significant moderation of time was found, b = 0.011, p = .159, 95% CI [-0.0043, 0.0264]. The regression test with the time and SE as predictors showed no funnel plot asymmetry, Z = -0.68, p = .495. These results indicate that moderation of the average ostracism effect is *not* found at a later time point in the included studies and time itself does not moderate the computed interaction effects. However, statistical sensitivity analyses showed that this interaction *was* significant when we removed three outliers based on studentized residuals, $\Delta d_{intercept}$ = -0.32, p = .029, 95% CI [-0.60, -0.03], whereas the regression coefficient time continued to be non-significant, b = 0.0002, p = .207, 95% CI [-0.0001, 0.0006]. On the last measure, this indicates that the non-significant interaction effect is sensitive to outliers in the data.

To see whether the interaction effects changed from the first to the last measure, we again compared confidence intervals. On the first measure, the 95% CI was [-0.64, -0.28] whereas for the last measure, the 95% CI was [-0.32, 0.05]. Considering the overlap of these CIs, one needs to be careful to interpret this as a reduction in the moderation across the measures examined. It is clear, however, that the average effect size of the interaction does not increase from first to last measure.

Secondary analyses

In addition to the simple effects over all studies, we analyzed subsets of studies that differ in type of dependent measure to study robustness of the effects. We also inspected whether sample composition, scale composition, and Cyberball specifics could predict the estimated effect size. Finally, we selected a homogeneous subset of studies to come to grips with the relatively large heterogeneity of simple main effects found for the primary hypotheses.

Measures

To inspect the robustness of the estimates of the first and last measure, we studied simple effects across several subsets of measures. These subsets encompassed interpersonal measures (i.e., measures that relate to others or the self in the context of others), intrapersonal measures (i.e., measures that relate only to the self), fundamental needs (single- and composite needs), and measures that were coded by the first two authors as fitting the description of being immediate or delayed (i.e., questions related to during- or after the game, respectively; shown in Fig. 3 as *model*). We ran the analyses for the different measures for the two time points separately (i.e., first and last measure).

Fig. 3. Dotplots of the average estimated simple effects with 95% confidence intervals. T1 represents first measure and T2 represents last measure. These effects are across the same subset. Traditional ostracism effect refers to the between-subjects effect of being ostracized with *no* moderator present, whereas moderated ostracism effect refers to being ostracized *with* a moderator present. Vice versa, moderator effect within ostracism/inclusion level refers to the between-subjects effect of the moderator factor, within the ostracized/inclusion conditions. The subset labeled “All” contains all measures. The subset labeled “Fundamental” contains only fundamental need measures. The subset labeled “Intrapersonal” contains all intrapersonal measures. The subset labeled “Interpersonal” contains all interpersonal measures. The subset labeled “Model” contains those where first measures is immediate and last measure is delayed. See Supplement S4.

The different panels in Fig. 3 show the results for the different simple effects per subset and overall; Table 3 summarizes the estimated interaction effects. A comparison of the results within each panel shows whether the overall results are robust and representative of all subsets, or whether there are nuances per type of measure. The main differences are notable in panels (1), (2), and (5). The first and second panels indicate that the effect of ostracism is weaker for interpersonal measures, compared to all intrapersonal measures (including fundamental needs). This indicates that in a similar factorial design, interpersonal measures show weaker effects than intrapersonal measures. Panel 5 indicates that the moderation of interpersonal measures is stronger compared to the other subsets. This suggests that interpersonal measures are more subject to moderation, whereas the effects of ostracism on interpersonal measures are smaller initially. Additionally, for the specific subset of fundamental needs, we noted that the point estimated interactions (Table 3) follow the pattern predicted by the need-threat model [11]: the first measures are moderated less strongly than the last measures.[5]

Table 3. Interaction effect per subset.

		k	Estimate	(*SE*)	*Z*-value	*p*-value	95% CI Lowerbound	95% CI Upperbound
Overall	T1	52	-0.46	0.09	-5.08	< .001	-0.64	-0.28
	T2	46	-0.19	0.11	-1.82	.069	-0.40	0.02
Fundamental	T1	30	-0.39	0.12	-3.42	< .001	-0.62	-0.17
	T2	17	-0.77	0.25	-3.05	.002	-1.27	-0.28
Intrapersonal	T1	42	-0.31	0.09	-3.38	< .001	-0.49	-0.13
	T2	39	-0.21	0.11	-1.87	.062	-0.44	0.01
Interpersonal	T1	10	-1.03	0.18	-5.69	<.0001	-1.38	-0.67
	$T1_{listwise}$	6	-0.36	0.22	-1.63	.104	-0.79	0.07
	T2	6	0.63	0.62	1.02	.309	-0.58	1.84

Model	T1	36	-0.29	0.10	-2.99	.003	-0.48	-0.10
	T2	23	0.01	0.17	0.08	.938	-0.31	0.34

The subset labeled “All” contains all measures. The subset labeled “Fundamental” contains only fundamental need measures. The subset labeled “Intrapersonal” contains all intrapersonal measures. The subset labeled “Interpersonal” contains all interpersonal measures. The subset labeled “Model” contains those where first measures is immediate and last measure is delayed. See Supplement S4. Listwise deletion ensures that estimates are made on full rows in the data. Listwise deletion was applied in all the subsets, which only altered results for interpersonal measures.

Composition

To inspect for structural and sampling effects of the studies, we ran mixed-effect models on the 120 ostracism effects, on both the first and the last measure. Due to listwise deletion, only 45 of 120 effect sizes remained on the first measure and 41 of 95 effect sizes for the last measure. The predictors in the mixed effects model were (1) country (US, other Western country, Asian, other), (2) proportion of males in the study, (3) mean age of the sample, (4) number of players in the game, (5) length of the game (≤ 5min, 5-10 min or > 10 min), (6) the number of throws in the game and (7) type of needs scale referenced (by assigning unique values for every unique reference).

On the first measure, this model (k = 45) showed clear residual heterogeneity after controlling for these structural- and sampling aspects of the studies, Q_E (33) = 449.52, p < .001, estimated τ_{res}^2 = 0.90, 95% CI [0.54, 1.59], but no overall moderation, Q_M (11) = 10.75, p = .465. The different types of need scales [11,21,22] did not significantly moderate effect sizes, showing psychometric convergence among the three scales. Inspecting the predictors individually also showed no indication for moderation (ps > .137; see Table 4).

Table 4. Meta regression coefficients for composition effects (first measure; k = 45).

	Estimate	(*SE*)	*Z*-value	*p*-value	95% CI Lowerbound	95% CI Upperbound
Intercept	-2.14	3.27	-1.89	0.058	-4.35	0.07
Structural						
Nr. of players	-0.22	1.05	-0.21	0.837	-2.28	1.85
Nr. of throws	0.03	0.02	1.49	0.137	-0.01	0.07
Ostracism <5 min	-	-	-	-	-	-
Ostracism 5-10 min	0.75	0.81	0.92	0.358	-0.84	2.34
Need scale = Williams (2000)	-	-	-	-	-	-
Need scale = Zadro et al. (2004)	-0.36	0.41	-0.88	0.381	-1.16	0.45
Need scale = Van Beest & Williams (2006)	0.07	0.54	0.13	0.894	-0.98	1.12
Need scale = Williams Zadro	-0.03	0.62	-0.04	0.965	-1.25	1.19
Need scale = Gonsalkorale & Williams (2007)	0.68	0.82	0.82	0.414	-0.94	2.30
Sampling						
Country = US	-	-	-	-	-	-
Country = Western	-0.42	0.36	-1.15	0.249	-1.13	0.29
Country = Asian	-0.30	1.13	-0.26	0.793	-2.51	1.92
Proportion male	1.54	1.09	1.42	0.156	-0.59	3.68
Mean age	-0.05	0.05	-0.97	0.332	-0.16	0.05

This can be interpreted as a standard regression formula. Empty rows represent reference categories.

On the last measure (*k* = 41; Table 5), no overall moderation was found, Q_M (11) = 6.00, *p* = .873, but heterogeneity did occur, Q_E (29) = 214.69, *p* < .0001. The number of players in the game significantly predicted the effects, *b* = 1.55, *p* = .047, 95% CI [0.2; 3.07], which would be interpreted as four players eliciting smaller ostracism effects, when compared to three players. The significance of this individual predictor should be interpreted carefully, as the omnibus moderation test showed no systematic decrease in heterogeneity. Overall, we found no strong evidence for moderation due to study or sample composition.[6]

Table 5. Meta-regression coefficients for composition effects (last measure; k = 41).

	Estimate	(*SE*)	*Z*-value	*p*-value	95% CI	95% CI
					Lowerbound	Upperbound
Intercept	-1.12	0.92	-1.21	0.227	-2.95	-0.70
Structural						
Nr. of players	1.55	0.78	1.98	0.047	0.02	3.07
Nr. of throws	0.01	0.02	0.59	0.556	-0.02	0.04
Ostracism <5 min	-	-	-	-	-	-
Ostracism 5-10 min	0.38	0.62	0.61	0.539	-0.83	1.59
Need scale = Williams (2000)	-	-	-	-	-	-
Need scale = Zadro et al. (2004)	-0.14	0.32	-0.44	0.658	-0.77	0.49
Need scale = Van Beest & Williams (2006)	-0.21	0.41	-0.51	0.613	-1.02	0.60
Need scale = Williams Zadro	-0.12	0.53	-0.22	0.826	-1.16	0.92
Need scale = Gonsalkorale & Williams (2007)	-0.07	0.65	-0.10	0.916	-1.33	1.20
Sampling						
Country = US	-	-	-	-	-	-
Country = Western	0.26	0.30	0.87	0.387	-0.33	0.86
Country = Asian	0.85	0.84	1.01	0.313	-0.80	2.49

Proportion male	0.29	0.83	0.35	0.730	-1.34	1.91
Mean age	-0.01	0.04	-0.25	0.806	-0.10	0.08

This can be interpreted as a standard regression formula. Empty rows represent reference categories.

Homogeneity?

The analysis of the simple ostracism effect on the first measure showed that differences of underlying effects made up 93% of the variability in study outcomes. We performed an additional secondary analysis in a more homogenous subset of studies to better understand this heterogeneity. This subset only included typical Cyberball studies that involved three players in the game, 30 throws, and lasted less than five minutes. In addition, the homogeneous subset of typical Cyberball studies only involved measures of immediate fundamental needs (single or composite). Performing a meta-analysis on this homogeneous subset of 19 studies showed an I^2 value of 83%, indicating that 83% of the total variability can be attributed to heterogeneity in the effect sizes. We noted that the mean simple ostracism effect in these 19 studies was relatively strong and estimated at d = -2.05, 95% CI [-2.44, -1.65]. In other words, given that the heterogeneity remains large even in a homogeneous subset, suggests that the heterogeneity found in the overall analyses does not appear to be an artifact from the inclusion of different measures and the use of alternative Cyberball setups.

Discussion

In this meta-analysis of Cyberball studies we estimated the average ostracism effect of the first and last dependent variable used in 120 Cyberball experiments. The primary hypotheses were (a) that the ostracism effect size would decrease from first to last measure and (b) that first measures would be less affected by cross-cutting variables than last measures. The secondary hypotheses tested whether the above generalizes across structural variables of the game, sample characteristics, or type of dependent variable used.

The results confirmed the hypothesis that the ostracism effect decreased from the first (d = -1.36) to the last measure (d = -.76), although this decline was not predicted by our estimation of duration between first and last measure. The results did not fully confirm the hypothesis that last measures are more strongly moderated than first measures. That is, our analysis of the experiments that included an experimentally controlled cross-cutting variable revealed that cross-cutting variables moderated both the first and last measure. In fact, visual inspection of the average estimated interaction effect sizes actually decreased in size from first (Δd = -.46) to last (Δd = -.19), although confidence intervals of these estimates did overlap.

To interpret the interactions it is important to recall (see Fig. 3) that the *overall* ostracism effects are relatively large and operated similarly at both levels of the cross-cutting moderator variable. Moreover, when we compared the mean effects of the moderator variable *within* the two possible levels of ostracism factor (i.e., ostracized or include), results indicate a relatively weak *positive* effect within the ostracism level and a relatively weak *negative* effect within the inclusion level. To further explain the implication of the findings it may be fruitful to consider an example in which participants are ostracized or included by either an outgroup or an ingroup. In such a setting, our findings would thus suggest that the relative effect of ostracism compared to inclusion (i.e., the ostracism effect), is similar for both outgroup *and* ingroup conditions. Moreover, if one compares the effect of group status (outgroup vs. ingroup), one would predict that those ostracized by outgroup members would slightly benefit whereas those included by ingroup members would slightly be harmed. Taken together, these contrasts support the robustness of the ostracism effect.[7]

Structural Aspects of Cyberball and Different Dependent Variables

The secondary analyses confirmed that the overall findings generalize to a large extent across structural aspects, sampling aspects and type of dependent variable.

Does gender of participants matter?

Previous research provided evidence for a difference in the ostracism effect across genders [18]. Our results indicated that, contrary to this, proportions of males and females did not significantly predict the mean effect size. In our coded studies, the mean proportion of males was approximately 39% (observed range: 0-100%).

Does age of participants matter?

Whereas previous research has indicated increased sensitivity to ostracism in younger age groups [19], we failed to find moderation of ostracism effects by mean age of the study samples. Coded studies had a mean sample age ranging from 10 through 32.5 years, with an average of approximately 20.5 years. This indicates that most of the research with Cyberball has been done on young adults, with relatively few or no studies investigating children, middle-aged participants, or senior citizens. More research could focus on specific (individual-level) age moderation of ostracism.

Does culture or country matter?

We found no indication that culture predicted the average effect size. In our coded studies, approximately 52% were from the United States, 45% from other Western countries (e.g., Australia, the Netherlands, Germany), and 3% from Asian countries. Our analyses used the United States as reference category. We note that the low prevalence of Asian countries might cause a lack of power and that we cannot definitively state there is no difference between Western and Asian responses to ostracism. We can state that there is no systematic difference in the ostracism response for Western countries and the United States.

Does number of players matter?

In the studies included in this meta-analysis, approximately 89% of the studies used the three-player version of Cyberball and 11% used the four-player version of Cyberball. Average ostracism effects differed between these subsets, with smaller predicted effects in the four-player setting, but we are hesitant to interpret this due to a nonsignificant omnibus test for the predictive model (see 'Composition' in the results section). Preferably, this moderator of the ostracism effect in Cyberball should be subject to further work in which the number of players is experimentally varied.

Does number of throws or length of the study matter?

We considered the length of Cyberball in two ways. We coded the number of ball tosses and estimated the length of the study. Of the coded studies, 60% used 30 throws, 11% used 40 throws, 8% used 20 throws, 4% used 60 throws, and 2% for both 15 and 24 throws. Other categories ranging from 10 through 200 make up the remaining percentages, each making up 1%. Only 2 out of 120 studies were estimated to last longer than 5 minutes. Our results indicated the mean ostracism effect was *not* reliably predicted to be different across different lengths of the study or the different number of total throws in the omnibus test. The single meta-regression on ball tosses suggested it may predict the effect size of the first measure. As above, we are hesitant to interpret this, but do note that increasing ball tosses may be more associated with a diffused ostracism effect than with an increased ostracism effect.

Does type of dependent variable matter?

Secondary analyses also showed that the majority of the results were robust across subsets of dependent measures and the overall set of dependent measures (see Fig. 3). Exceptions were interpersonal measures showing relatively weaker ostracism effects on the first measure when compared to the other subsets. This suggests that psychological effects of ostracism are large, but that this effect might be smaller for interpersonal behaviors. On top of this, interpersonal measures also show more moderation, suggesting that interpersonal behaviors caused by ostracism are more easily moderated by cross-cutting factors. Additionally, we estimated interactions for the measure subsets interpersonal (i.e., measures relating to others), intrapersonal (measures relating to the self), fundamental needs, model (i.e., first measure is reflexive and last measure is reflective), and an overlap of the latter two subsets. For all but two, these subsets showed that measures taken at the first time point were moderated more strongly than the measures taken last. Finally, the analyses including only fundamental needs showed that moderation was larger at the last time point, when compared to the first time point. This result is crucial, as Williams [11] specifically predicted this pattern for fundamental needs.

Williams's Model of Ostracism: Supported or Not?

Regarding the test of Williams's [11] model, there are several important observations and limitations. First, Williams proposed fundamental need threat as a result of even a brief episode of ostracism. This was supported by the meta-analysis. Moreover, moderation is predicted to occur in the reflective stage, when the context and meaning of the ostracism event can be appraised. This was also supported in the present meta-analysis. The final stage of Williams's model—resignation—is outside the aims of the present meta-analysis, because it requires long-term exposure to ostracism.

The proposition that appears to lack support from this meta-analysis is that reflexive reactions to ostracism are more resistant to moderation than reflective reactions. Across the board, our results indicate there is more moderation of ostracism effects on the first time point than on the last time point. However, there are two limitations to this conclusion. Firstly, Williams specifically refers to physiological, online, or immediate retrospective reports to assess reflexive reactions. In many instances in this meta-analysis, the first reaction is not isomorphic with reflexive measures. Anything taken after the game, or assessed by wording indicating present state (rather than the participants' state during the game), is not assumed to be reflexive, nor predicted to be resistant to moderation. Secondly, Williams's proposition is restricted to fundamental needs only. Indeed, our specific analyses involving only studies that employed measures of immediate and delayed fundamental need satisfaction corroborated the model prediction that there is more moderation on the last time point, than on the first time point.

Because of this quantitative difference in moderation across measures, we encourage direct testing of this time difference in moderation as predicted by Williams [11], just as the study by Bernstein and Claypool [39] was a direct, experimental test of a finding by Gerber and Wheeler [14]. However, the mean size of the interaction effect in out meta-analysis was quite small, raising power issues for future studies. Using our estimated interaction effects to determine sample size under a power of .8, a sample size of 2186 would be necessary to have sufficient power on both time points.[8] Note that the mean sample size in full factorial designs in our meta-analysis is 110, showing that the mean power in these studies is .08 to detect an *interaction* at the last time point (notably, power for the standard ostracism effect is highly sufficient in the included studies, due to the large effect). A large Mechanical Turk study is feasible and could provide the sample needed. Additional ways of increasing power are by reducing error on the measurements by using validated psychometric scales.

Changes to the need-threat model of ostracism

As a result of our findings, we suggest that the temporal need-threat model of ostracism should be modified. Firstly, it should be recognized that there is potential for moderation in the reflexive stage, where immediate measures of impact tap into participants' reactions during the game. If factors can reduce physical pain and distress, like for instance acetaminophen [40][9] or transcranial magnetic stimulation [41], or if certain populations are less likely to feel pain (e.g., those higher in schizotypal personality disorder [42,43]), then we would also expect moderation of immediate measures of distress. Secondly, our results may suggest important issues related to the timing of measuring ostracism effects by way of the ordinal differences. Specifically, time passed after the ostracism episode occurred is likely to affect the extent immediate distress measures will be subject to moderation. For example, if researchers wait long enough before administering the immediate need satisfaction measures (e.g., “playing the game made me feel insecure”), it becomes more likely that all participants will have recovered from the negative impact of ostracism, thus resulting in a homogeneous (and highly satisfied) between-group result. Thus, differences in recovery from ostracism based upon social-situational factors and/or personality differences, if any, occur somewhere between initial pain and final recovery. It is difficult to predict exactly when that time period is. Zadro et al. [44] report delayed recovery by those high in social anxiety 45-minutes later. Other studies show full recovery within 5-10 minutes. Future research needs to examine the time course more carefully, to determine if and when moderation occurs in delayed measures.

Limitations

Within the current meta-analysis there are several limitations. One potential limitation is that our testing of differences between first and last measure was indirect. We compared confidence intervals to evaluate whether the effects were different. A direct test would provide more conclusive evidence on whether or not the effects are indeed equal or different across the first and last measurements. Note, however, that a direct test requires correlations between the measurements for every study, every condition, and every type of different variable. This information was not given in the vast majority of the papers and we anticipated that a direct request for such information would suffer from the problem of low response rates [45] which would in turn lower the sample size of the meta-analysis and thus the ability to effectively test our hypotheses.

A second potential limitation is that the random (non-systematic) heterogeneity in the effect sizes poses a problem for the power of finding moderator effects [25]. This could pose the problem that several of the non-effects found are actually there, but not detected (Type II errors). However, our subset analysis of typical Cyberball studies – 3 players games involving 30 ball tosses, lasting less than five minutes, with immediate fundamental need satisfaction as dependent variable - still showed substantial variability in the effect sizes: I^2 = 83%. This indicates that the effects are quite variable to begin with and makes it unlikely that the overall effects are misrepresented.

Also, we did not observe that our estimation of time predicted the ostracism effect on the last measure. This null-effect may be a reality but could also be caused by the fact that the (random) heterogeneity in the effect sizes may have been too large to find moderation by time. This cannot be counteracted in the current dataset and remains a limitation. Second, imprecise reporting of the measures in the papers may have led to inaccurate time estimations. To counteract this imprecise reporting of measures, authors could be contacted, but this also poses new problems (i.e., nonresponse, or authors might not be willing to admit that measures were left out in the paper [46]).

Importantly, we did observe that the confidence intervals of both the first and last measure did not overlap, suggesting that there is a difference in effect size between first and last measure. The question then is whether this difference is indeed caused by time of measurement or in part caused by the type of measurement used across the two different time points. This explanation can be addressed by inspecting whether the composition of measures is different across time points. On the first measure 0.84 was intrapersonal self-report, 0.02 was intrapersonal physiological, 0.01 was intrapersonal other, 0.08 was interpersonal anti-social, 0.03 was interpersonal pro-social, and 0.01 interpersonal other. On the last measure 0.79 was intrapersonal self-report, 0.04 was intrapersonal physiological, 0.02 was intrapersonal other, 0.05 was interpersonal anti-social, 0.08 was interpersonal pro-social, and 0.01 was interpersonal other. This shows that the different types of dependent variables are similarly distributed across time points (maximum discrepancy of 4.9 percentage points). Substantive differences in proportions of measures across time points are minimal and thus form an unlikely driving force for our findings.

A third limitation is that this paper only summarized the results of the measures included in the studies. However obvious this might be, it should be pointed out, because the validity of the conclusions are reliant on the validity of the measures. Most prominently represented in the current meta-analysis are the fundamental need measures, which have no proper psychometric validation up-to-date, notwithstanding their wide use. Other kinds of included measures possibly also lack proper validation and one has been openly criticized (e.g., the Hot Sauce aggression paradigm [47]).

Conclusion

Our meta-analysis of 120 Cyberball studies extends the temporal need-threat model of ostracism. We observed that the average effect size approaches 1.5 standard deviations and that this average effect size is not affected by the composition of the sample used (i.e., age, gender, country of origin) nor by structural aspects of the game (i.e., number of ball tosses, duration, players). We also observed that findings are relatively robust across the typical dependent variables that are used in Cyberball and that the overall effect size decreases from first to last measure. Importantly, we also observed that first measures can be moderated by cross-cutting variables and that only fundamental needs measures show stronger moderation for the last measures as opposed to the first measure taken in the studies. The moderation analyses by cross-cutting variables also revealed that the interaction effects sizes are considerably smaller than the direct inclusion vs. ostracism effect size. This revealed that the typical Cyberball study has enough power to detect main effects, but should substantially increase sample size to study theoretically relevant interactions. Intriguingly, we also observed that effect sizes were rather heterogeneous even when we limited our analysis to a very homogenous subset of studies. This indicates that there are potentially relevant moderators that have yet not been discovered. We invite fellow researchers to reanalyze our data (osf.io/ht25n) and test new hypotheses, and to further expand our knowledge of ostracism with Cyberball.

Footnotes

1. The direct link: https://osf.io/ht25n/
2. It has been updated since, but the list that was used can be found on the Open Science Framework, see Footnote 1.
3. Oaten, Williams, Jones and Zadro [48] was applicable, but was excluded due to being an outlier with respect to effect size (*d*s > 15). See also Gerber and Wheeler (2009; p. 473): "*One study (Oaten, Williams, Jones, & Zadro, 2007) had need effect sizes that were clear outliers (effect sizes were 5–7 standard deviations above the means)* […and…] *were excluded from the analyses.*"
4. Due to the dependency between the standardized effect size and the standard error, we also ran an alternative version of the Egger's test that regresses on 1/N. These analyses yielded highly similar results.
5. Because fundamental needs showed effects in the theorized direction, we explored this further by overlapping the subset of fundamental need measures with the model definition of immediate and delayed (i.e., whether the measures related to feelings during or after the Cyberball game). Estimated interactions for this selection were Δd = -0.37, 95% CI [-0.60, -0,14] ($k = 29$) and Δd = -0.13, 95% CI [-0.53, 0.27] ($k = 8$) for the first and last measure, respectively. So in this particular subset of studies that use immediate or delayed fundamental needs measures, results are not in line with Williams's [11] prediction. The reported fundamental need selection can be specified even further to only include studies that explicitly focus on composite need satisfaction as typically defined by Kip Williams. Such a selection again provides support for the hypothesis that immediate fundamental need satisfaction is less moderated, Δd = -0.18, 95% CI [-0.47, -0.11] ($k = 15$), than delayed need satisfaction,

$\Delta\underline{d}$ = -0.93, 95% CI [-1.67, -0.19] (k = 3). Note, however, that such a selection is based on 3 studies for delayed measures.

6. We also conducted individual meta-regressions for each of the structural- and sampling variables. These individual analyses yield similar results as the overall analyses. We again observed that four players are less hurt by ostracism than three players (b = .84, SE = .28, p = .003) on the last measure. What is new is that we also observed that number of ball tosses affected the effect size (b = .02, SE = .01, p = .046) on the first measure. This showed that increasing the number of ball tosses decreases the negative impact of ostracism. Taken together this suggests that the impact of ostracism is diffused when it is the result of more players and more ball tosses compared to fewer players and fewer balls tosses.
7. It is important that the simple effects in Fig. 3 are averaged over studies, thus potentially subject to Simpson's paradox.
8. We used G*Power 3.1.7 to calculate this between-subjects interaction effect (F-test, fixed effects, .8 power); with k = 4 and the smaller interaction (last time point; numerator $df = k – 1$). The effect size Δd was transformed in to f by means of $\surd[d^2/(2k)]$, resulting in f = .0707.
9. DeWall et al. was not included in the meta-analysis, because we were not able to retrieve all information.

Appendix

All formulae reported below originate from the chapter by Michael Borenstein (2009). Hedges' *g* was calculated as

$$g = d\left(1 - \frac{3}{4df_w - 9}\right)$$

where *d* is the standardized main effect and df_w is the number of conditions minus 1. For the standardized interaction effect *d* was calculated as

$$\Delta d = \frac{(\bar{X}_{11} - \bar{X}_{12}) - (\bar{X}_{21} - \bar{X}_{22})}{s_p}$$

where the first term in the numerator is the ostracism effect and the second term is the ostracism effect in the moderator conditions. When transformed to a squared correlation coefficient, this Δ*d* corresponds to the partial eta-squared of the interaction. Sampling variance of *g* was calculated by multiplying the sampling variance of *d* by the squared correction factor, that is

$$s_g^2 = \left(1 - \frac{3}{4df_w - 9}\right)^2 s_d^2$$

where the sampling variance of the interaction was calculated as the sum of the sampling variances of both the simple main effects.

Acknowledgements

We would like to kindly thank all researchers who sent in their unpublished studies and helped us collect all the data we needed. We would also like to thank Andrew Hales for his constructive comments and feedback. The preparation of this article was supported by grant number 016-125-385 from the Netherlands Organization for Scientific Research (NWO) awarded to Jelte M. Wicherts and by the NSF under Grant #BCS-1339160 awarded to Kipling D. Williams.

References

References marked with an asterisk indicate studies included in the meta-analysis.

1. *Williams KD, Cheung CK, Choi W (2000) Cyberostracism: effects of being ignored over the Internet. J Pers Soc Psychol 79: 748–762.
2. Baumeister RF, Twenge JM, Nuss CK (2002) Effects of social exclusion on cognitive processes: Anticipated aloneness reduces intelligent thought. J Pers Soc Psychol 83: 817–827.
3. Nezlek JB, Kowalski RM, Leary MR, Blevins T, Holgate S (1997) Personality moderators of reactions to interpersonal rejection: Depression and trait self-esteem. Personal Soc Psychol Bull 23: 1235–1244.
4. Craighead WE, Kimball WH, Rehak PJ (1979) Mood changes, physiological responses, and self-statements during social rejection imagery. J Consult Clin Psychol 47: 385–396.
5. Leary MR, Kowalski RM, Smith L, Phillips S (2003) Teasing, rejection, and violence: Case studies of the school shootings. Aggress Behav 29: 202–214.

6. *Lustenberger DE, Jagacinski CM (2010) Exploring the Effects of Ostracism on Performance and Intrinsic Motivation. Hum Perform 23: 283–304.

7. *Carter-Sowell AR, Chen Z, Williams KD (2008) Ostracism increases social susceptibility. Soc Influ 3: 143–153.

8. *Van Beest I, Carter-Sowell AR, van Dijk E, Williams KD (2012) Groups being ostracized by groups: Is the pain shared, is recovery quicker, and are groups more likely to be aggressive? Gr Dyn Theory, Res Pract 16: 241–254.

9. Baumeister RF, Leary MR (1995) The need to belong: desire for interpersonal attachments as a fundamental human motivation. Psychol Bull 117: 497–529.

10. *Ijzerman H, Gallucci M, Pouw WTJL, Weiβgerber SC, Van Doesum NJ, et al. (2012) Cold-blooded loneliness: social exclusion leads to lower skin temperatures. Acta Psychol (Amst) 140: 283–288.

11. Williams KD (2009) Ostracism: a temporal need-threat model. Adv Exp Soc Psychol 41: 275–314.

12. Haselton MG, Buss DM (2000) Error management theory: a new perspective on biases in cross-sex mind reading. J Pers Soc Psychol 78: 81–91.

13. Blackhart GC, Nelson BC, Knowles ML, Baumeister RF (2009) Rejection elicits emotional reactions but neither causes immediate distress nor lowers self-esteem: a meta-analytic review of 192 studies on social exclusion. Pers Soc Psychol Rev 13: 269–309.

14. Gerber J, Wheeler L (2009) On Being Rejected: A Meta-Analysis of Experimental Research on Rejection. Perspect Psychol Sci 4: 468–488.

15. Cacioppo S, Frum C, Asp E, Weiss RM, Lewis JW, et al. (2013) A Quantitative Meta-Analysis of Functional Imaging Studies of Social Rejection. Sci Rep 3.

16. Rotge J-Y, Lemogne C, Hinfray S, Huguet P, Grynszpan O, et al. (2014) A meta-analysis of the anterior cingulate contribution to social pain. Soc Cogn Affect Neurosci: nsu110.

17. *De Waal-Andrews W, van Beest I (2012) When you don't quite get what you want: psychological and interpersonal consequences of claiming inclusion. Pers Soc Psychol Bull 38: 1367–1377.

18. *Hawes DJ, Zadro L, Fink E, Richardson R, O'Moore K, et al. (2012) The effects of peer ostracism on children's cognitive processes. Eur J Dev Psychol 9: 599–613.

19. *Pharo H, Gross J, Richardson R, Hayne H (2011) Age-related changes in the effect of ostracism. Soc Influ 6: 22–38.

20. Hofstede G (1980) Culture's consequences: International differences in work-related values. London, UK: Sage.

21. *Van Beest I, Williams KD (2006) When inclusion costs and ostracism pays, ostracism still hurts. J Pers Soc Psychol 91: 918–928.

22. *Zadro L, Williams KD, Richardson R (2004) How low can you go? Ostracism by a computer is sufficient to lower self-reported levels of belonging, control, self-esteem, and meaningful existence. J Exp Soc Psychol 40: 560–567.

23. Hunter J, Schmidt F (1990) Dichotomization of continuous variables: The implications for meta-analysis. J Appl Psychol 75: 334–349.

24. MacCallum RC, Zhang S, Preacher KJ, Rucker DD (2002) On the practice of dichotomization of quantitative variables. Psychol Methods 7: 19–40.

25. Hedges L V, Pigott TD (2004) The power of statistical tests for moderators in meta-analysis. Psychol Methods 9: 426–445.

26. Williams KD, Jarvis B (2006) Cyberball: A program for use in research on interpersonal ostracism and acceptance. Behav Res Methods 38: 174–180.

27. Smits IAM, Dolan C V, Vorst H, Wicherts JM, Timmerman ME (2011) Cohort differences in Big Five personality factors over a period of 25 years. J Pers Soc Psychol 100: 1124–1138.

28. *Gonsalkorale K, Williams KD (2007) The KKK won't let me play: ostracism even by a despised outgroup hurts. Eur J Soc Psychol 37: 1176–1186.

29. Viechtbauer W (2010) Conducting meta-analyses in R with the metafor package. J Stat Softw 36: 1–48.

30. R Core Team (2013) R: A language and environment for statistical computing. Available: http://www.r-project.org/.

31. Hedges LV (1981) Distribution theory for Glass's estimator of effect size and related estimators. 6: 107–128.

32. Viechtbauer W (2005) Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. J Educ Behav Stat 30: 261–293.

33. Light RJ, Pillemer DB (1984) Summing up: the science of reviewing research. Cambridge, MA: Harvard University Press.

34. Bakker M, Van Dijk A, Wicherts JM (2012) The rules of the game called psychological science. Perspect Psychol Sci 7: 543–554.

35. Egger M, Smith GD, Schneider M, Minder C (1997) Bias in meta-analysis detected by a simple, graphical test. BMJ 315: 629–634.

36. Sterne JAC, Egger M (2005) Regression Methods to Detect Publication and Other Bias in Meta-Analysis. In: Rothstein HR, Sutton AJ, Borenstein M, editors. Publication bias in meta-analysis. Chichester: John Wiley & Sons.

37. Cohen J (1988) Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum.

38. Schenker N, Gentleman JF (2001) On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals. Am Stat 55: 182–186.

39. *Bernstein MJ, Claypool HM (2012) Not all social exclusions are created equal: Emotional distress following social exclusion is moderated by exclusion paradigm. Soc Influ 7: 113–130.

40. DeWall CN, MacDonald G, Webster GD, Masten CL, Baumeister RF, et al. (2010) Acetaminophen reduces social pain: behavioral and neural evidence. Psychol Sci 21: 931–937.

41. *Riva P, Romero Lauro LJ, Dewall CN, Bushman BJ (2012) Buffer the pain away: stimulating the right ventrolateral prefrontal cortex reduces pain following social exclusion. Psychol Sci 23: 1473–1475.

42. *Wirth JH, Lynam DR, Williams KD (2010) When social pain is not automatic: Personality disorder traits buffer ostracism's immediate negative impact. J Res Pers 44: 397–401.

43. Lautenbacher S, Krieg J-C (1994) Pain perception in psychiatric disorders: A review of the literature. J Psychiatr Res 28: 109–122.

44. *Zadro L, Boland C, Richardson R (2006) How long does it last? The persistence of the effects of ostracism in the socially anxious. J Exp Soc Psychol 42: 692–697.

45. Wicherts JM, Borsboom D, Kats J, Molenaar D (2006) The poor availability of psychological research data for reanalysis. Am Psychol 61: 726–728.

46. LeBel EP, Borsboom D, Giner-Sorolla R, Hasselman F, Peters KR, et al. (2013) PsychDisclosure.org: Grassroots Support for Reforming Reporting Standards in Psychology. Perspect Psychol Sci 8: 424–432.

47. Ritter D, Eslea M (2005) Hot Sauce, toy guns, and graffiti: A critical account of current laboratory aggression paradigms. Aggress Behav 31: 407–419.

48. Oaten M, Williams KD, Jones A, Zadro L (2008) The effects of ostracism on self-regulation in the socially anxious. J Soc Clin Psychol 27: 471–504.

49. Borenstein M (2009) Effect sizes for continuous data. In: Cooper H, Hedges L V., Valentine JC, editors. The handbook of research synthesis and meta-analysis. New York, NY: Russell Sage Foundation.

Supporting information

S1 File. Data package. Contains data and the R analysis script.

S2 File. Full reference list meta-analysis studies. Contains the full reference list of the studies included in the meta-analysis.

S3 File. Scatterplot of the effects in hypotheses 1 and 2 and estimated time.

S4 File. Figure 3 subset lists. Contains the lists of what studies that were in the meta-analysis are included in computing the effects for the different panels.

Figure 1

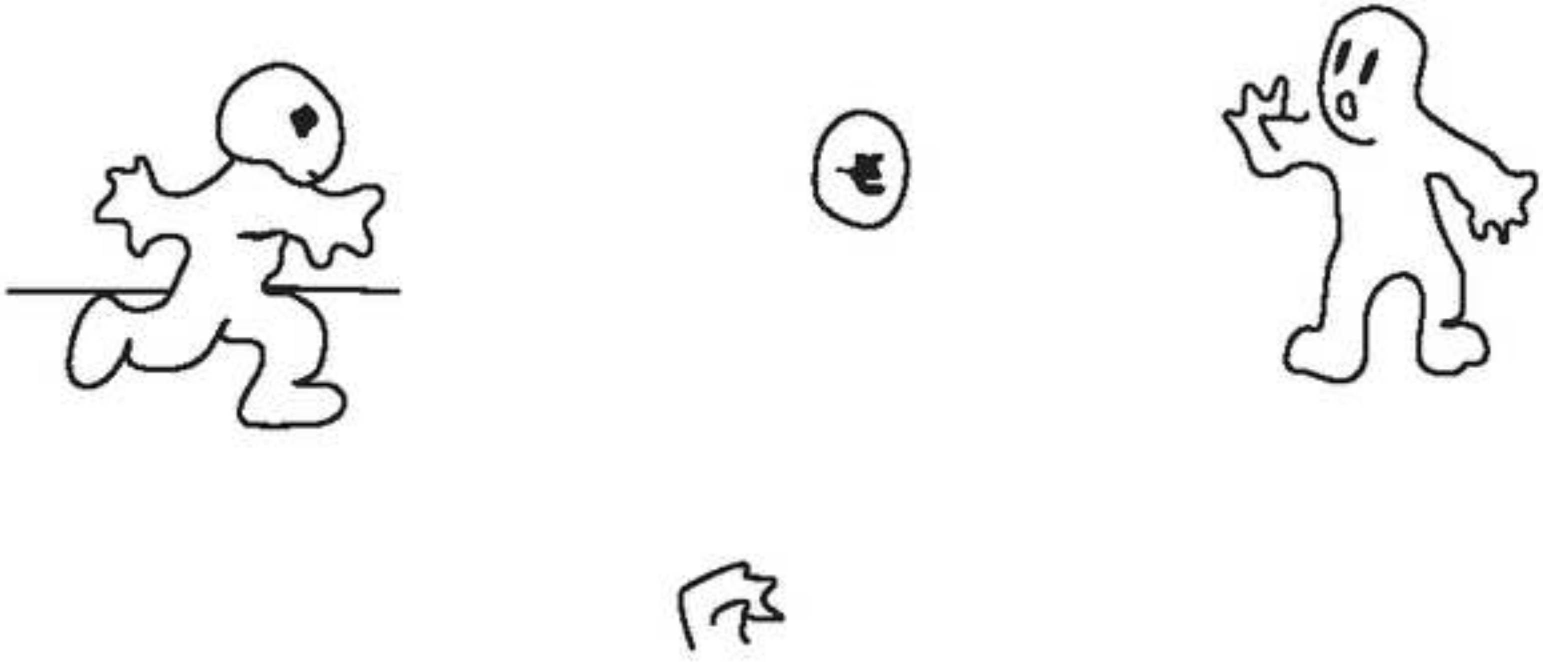


Figure 2

Click here to download Figure: Fig2.tiff

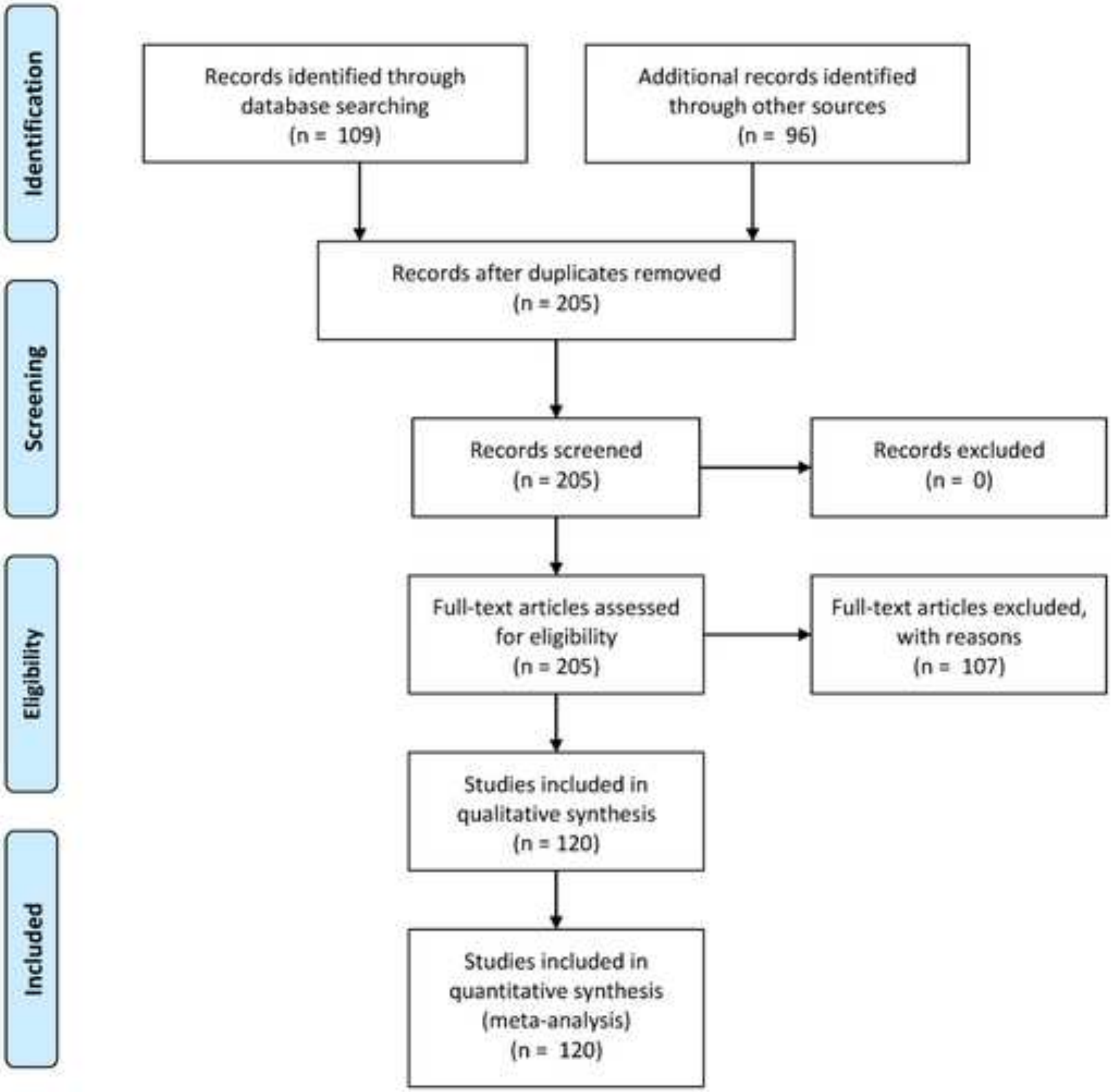
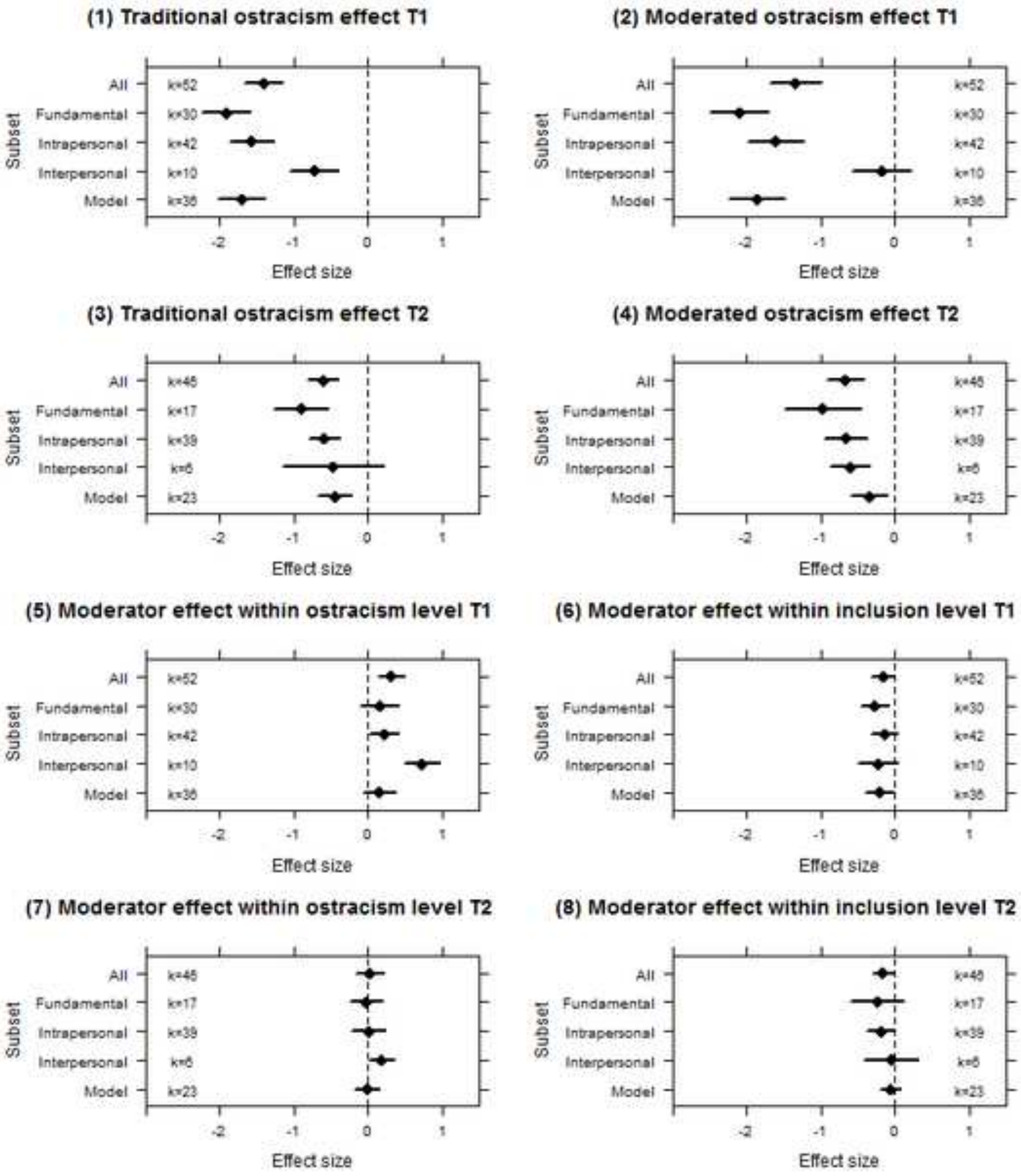

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 109)
Additional records identified through other sources (n = 96)
Records after duplicates removed (n = 205)
Records screened (n = 205)
Records excluded (n = 0)
Full-text articles assessed for eligibility (n = 205)
Full-text articles excluded, with reasons (n = 107)
Studies included in qualitative synthesis (n = 120)
Studies included in quantitative synthesis (meta-analysis) (n = 120)

Figure 3


(1) Traditional ostracism effect T1

(2) Moderated ostracism effect T1

(3) Traditional ostracism effect T2

(4) Moderated ostracism effect T2

(5) Moderator effect within ostracism level T1

(6) Moderator effect within inclusion level T1

(7) Moderator effect within ostracism level T2

(8) Moderator effect within inclusion level T2

Click here to download Supporting Information: PRISMA 2009 Checklist.doc

Click here to download Supporting Information: S3 Time effect scatterplot.tiff

The Ordinal Effects of Ostracism:

A Meta-Analysis of 120 Cyberball Studies

Chris H.J. Hartgerink[1] ¶

Ilja Van Beest[2*] ¶

Jelte M. Wicherts[1]

Kipling D. Williams[3]

[1] Department of Methodology and Statistics, Tilburg University, North-Brabant, the Netherlands

[2] Department of Social Psychology, Tilburg Univeristy, North-Brabant, the Netherlands

[3] Department of Psychology, Purdue University, Illinois, United States of America

*Corresponding author

E-mail: i.vanbeest@tilburguniversity.edu

¶ These authors contributed equally to this work.

Abstract

We examined 120 Cyberball studies (N = 11,869) to determine the effect size of ostracism and conditions under which the effect may be reversed, eliminated, or small. Our analyses showed that (1) the average ostracism effect is large (d > |1.4|) and (2) generalizes across structural aspects (number of players, ostracism duration, number of tosses, type of needs scale), sampling aspects (gender, age, country), and types of dependent measure (interpersonal, intrapersonal, fundamental needs). Further, we test Williams's (2009) proposition that the immediate impact of ostracism is resistant to moderation, but that moderation is more likely to be observed in delayed measures. Our findings suggest that (3) both first and last measures are susceptible to moderation~~,~~ and (4) time passed since being ostracized does not predict effect sizes of the last measure. Thus, support for this proposition is tenuous~~,~~ and we suggest modifications to the temporal need-threat model of ostracism.

Keywords: Cyberball, meta-analysis, ordinal, ostracism

Introduction

Cyberball [1] is a virtual ball-tossing game that is used to manipulate the degree of social inclusion or ostracism in social psychological experiments. In this game the participant supposedly plays with two (or more) other participants, who are in fact part of the computer program. The program varies the degree to which the participant is passed the ball (see Fig. 1 for a still from the game). Ostracized players are not passed the ball after two initial tosses and thus obtain fewer ball tosses than the other players. Included players are repeatedly passed the ball and obtain an equal number of ball tosses as the other players. Our literature search showed that at least 200 published papers involved the use of the Cyberball paradigm to study ostracism and that over 19,500 participants have played the game thus far. In this paper we provide a meta-analysis of these studies. Our aim was to gauge the typical effect size of being ostracized in the Cyberball game and to see whether this effect is moderated by cross-cutting variables that were hypothesized to reduce/enhance the psychological impact of ostracism, structural aspects that are inherent in Cyberball (e.g., number of players, number of ball tosses), sampling aspects of the studies (e.g., gender composition), the type of dependent variables used (e.g., intrapersonal measures such as need satisfaction or interpersonal measures such as pro- or antisocial behavior), and ~~(4)~~ the ordinal time point of the variable assessment (i.e., first or last).

Fig. 1. Cyberball game screenshot.

Historical background

Cyberball was introduced in 2000 as a means to study ostracism, that is:~~— i.e.,~~ being excluded and ignored [1].~~[1].~~ This focus of Cyberball on ostracism ~~makes it an unique~~

~~paradigm that~~ sets it apart from other paradigms that are tailored~~have been used~~ to study rejection, such as the future life rejection [2],~~[2],~~ the get-acquainted paradigm [3],~~[3],~~ and the autobiographical memory manipulation (i.e., remember a time when you were excluded [4]).~~[4]).~~ The difference is that participants in Cyberball are not explicitly informed that they are excluded whereas in the other paradigms participants are provided a reason pertaining to why they are excluded. ~~Cyberball participants simply do not obtain a ball and thus need to infer that they are excluded, whereas in the other paradigms, participants are informed that they are excluded in various ways and thus do not need to infer that they are excluded.~~

The Cyberball manipulation is a suitable method to study how people react to being ignored and excluded. Humans are social animals and care deeply about whether they are included or ostracized by others. Interestingly, ostracism is not only observed among loved ones, but on all levels of human organization. In fact, research suggests that most people are ignored and excluded at least once a day [3]. The social relevance is further evident in that ostracism~~it~~ not only affects the person who is ostracized (intrapersonal effects), but often also others (interpersonal effects). As a grim example, research on school shootings has suggested a direct link between ostracism and revenge. People~~, which does not only affect the people~~ who were ostracized may retaliate by murdering those responsible and sometimes even ~~, but also~~ innocent bystanders [5].~~[5].~~ The impact of ostracism is also evident in research findings using Cyberball. Through experimental work, it has been repeatedly shown that being ostracized has an effect on people—either on their psychological functioning (e.g., decreases in positive mood [6]) or on certain interpersonal behaviors (e.g., increases in social susceptibility or aggressive behaviors [7,8]). These experiments have highlighted the (mostly negative) impact of ostracism on fundamental needs (e.g., belonging [9]), mood, physiology (e.g., body temperature [10]), and various other constructs, including those measured with behavioral measures (e.g., conformity, compliance, aggression). In the current paper, we refer to the general effect of being ostracized compared to being included in Cyberball as the *ostracism effect*.

To capture how people respond to ostracism, Williams [11] proposed a temporal need-threat model of ostracism. Here he suggested three stages of the ostracism effect, namely: (1) a *reflexive* stage, (2) a *reflective* stage, and (3) a *resignation* stage. In the reflexive stage, the response to the ostracism sequence is immediate and occurs like a reflex. This initial response is theorized to be socially painful, threatening [9] and, following overdetection theory [12], should be~~This initial response is theorized to be socially painful, threatening [9] and~~ easily detectable due to evolutionary over-sensitivity to cues of ostracism.~~ [12].~~ Such a reflex would not take into account situational specifics and provides little room for coping. The reflex is proposed to affect primarily pain, fundamental needs, and emotional reactions (e.g., increased anger and sadness). The affected fundamental needs are belonging, self-esteem, control, and meaningful existence, typically measured by a need satisfaction scale [11]. According to Williams, measures of reflexive responses must occur during, or in the case of self-report measures, immediately following Cyberball (with the wording of the questions referring to how participants felt *during the game*). The *reflective* (or delayed) stage, which follows this immediate response, is subject to more rational thought and coping with the threats. Part of such coping is the necessity for fortification of the threatened fundamental needs. Coping can be measured both in terms of speed of recovery (higher levels of need satisfaction approaching the levels of included participants)~~),~~ and emotional, cognitive, and behavioral choices. The *resignation* stage occurs after prolonged ostracism, causing prolonged periods of pain and more fundamental need threat. If one is not able to fortify the fundamental needs, a prolonged ostracism sequence leads to feelings of helplessness, alienation, depression, and unworthiness. Because the resignation stage is hypothesized to occur only after prolonged and repeated exposure to ostracism (as in months or years), it is not feasible (and even unethical) to study resignation responses in laboratory experiments. Hence, in this paper we limit ourselves to studying the reflexive and reflective stages. For these stages, Williams asserts that moderation and variation of need satisfaction effects by individual differences and socially relevant factors (e.g., type of group from which one is excluded) will be less likely to occur for reflexive measures than for reflective measures.

Goals of meta-analysis

A limited number of Cyberball experiments have been reviewed in other meta-analyses, but these meta-analyses had a different goal than the current meta-analysis. Previous meta-analyses focused on social rejection and not on ostracism [13,14], or focused only on a specific dependent variable (e.g., fMRI [15,16]). Importantly, none of these early meta-analyses were specifically set up to test Cyberball effects only. Consequently, we do not know how structural variables of Cyberball or sample characteristics affect the ostracism effect size. Moreover, none of these meta-analyses considered whether it matters if a specific variable is measured first or last. Thus, it remains unclear whether the ostracism effect size decreases or increases over time and whether immediate measures are more or less moderated by cross-cutting variables. The goal of our meta-analysis is to provide a comprehensive understanding of the Cyberball-induced inclusion versus ostracism effect size. Under what conditions, if any, is the effect size negative, zero, or especially small? Under what conditions is it especially large? To answer these questions we made several selection decisions (see also the Open Science Framework (OSF) where we preregistered all selections and hypotheses).[1]

The first selection decision is that we considered only the first and the last dependent variable of all included studies. The reason for this selection was that it allowed us to gauge whether the effect sizes are affected by the time point at which the effects are measured. Another reason is that it served as a proxy to evaluate the hypothesis that immediate measures should be less affected by cross-cutting variables than more delayed measures.

A second decision is that we considered two different approaches to test whether first and last measures can be moderated by cross-cutting variables. This allowed us to test the robustness of our hypothesis across independent variables. The first approach to assess moderation was to conduct a meta-analysis~~analyses~~ on all studies that were explicitly designed to test whether being ostracized or included can be moderated by a cross-cutting factor. For this purpose we selected all the studies that included an experimentally manipulated moderator variable. Moreover, to meta-analyze the interaction term for first and last measure we followed the prediction of the authors in computing this interaction term. A potential limitation of our decision to follow the prediction of the authors is that the predictions may have been generated post-hoc on the basis of observed outcomes. For example, if authors used a 2 (ostracized vs included) x 2 (ingroup vs outgroup design) we followed the prediction of the authors to compute whether the interaction term denotes that ostracism is increased by an outgroup or decreased by an outgroup (specific calculations are reported in the methods section and formulae in the Appendix). Moreover, after computing the overall interaction terms we created dotplots in which we depicted the effect of ostracism across the two levels of the moderator~~,~~ and – perhaps more importantly - the effect of the moderator across the two levels of the ostracism manipulation. This was done to facilitate the interpretation of an interaction term and specifically to show whether cross-cutting variables have more impact on being included in Cyberball or more impact on being ostracized in Cyberball [17].

The second approach to test moderation was to assess if and how first and last measures are moderated by structural aspects of Cyberball (i.e., number of depicted Cyberball players, number of ball tosses used, duration of the game) and sample aspects (i.e., gender composition, country of origin, age). Note that the outcome of this analysis may thus also be used for future researchers to decide how to set ~~-~~up a game of Cyberball and whether effects generalize across age, gender, and country of origin. Because prior research has not explicitly manipulated structural aspects in controlled experiments we did not have a specific prediction whether increasing the number of players, ball tosses, and game duration would increase or diffuse the impact of ostracism. Given that the social aspects of an interdependent setting may be less evolutionary relevant for males than for females [18]~~,~~ and less relevant for older people than younger people [19], we explored whether an increase of male participants and mean age would decrease the ostracism effect. Moreover, considering that collectivism might influence the degree to which belonging is important [20], we used a categorization of continents (i.e., U.S., other western countries, Asian countries, and remaining countries) to explore whether a more collective orientation would be associated with larger ostracism effects. Finally, because some of the factors might be related (i.e., an increased number of ball tosses is likely to be associated with an increase in duration), we decided to use a regression approach in which all factors were entered simultaneously. A benefit of this approach is that it ensures that significant predictors have an impact above and beyond the impact of the other predictors.

The third decision is that we also checked the robustness of our findings across various dependent variables. More specifically, we coded whether the first and last measures belonged to the category of *interpersonal* variables assessing how ostracism impacts others or belonged to the category of *intrapersonal* variables assessing how ostracism impacts the self. Examples of interpersonal measures are donations to charity, helping behavior, money allocations in economic games, and aggression measures such as irritating sounds blasts or hot sauce allocation. These were initially coded into pro- and anti-social, but were collated into the category interpersonal due to small k the first measure (4 and 10, respectively) and last measure (8 and 6, respectively). Examples of intrapersonal measures are self-reported anger, self-esteem, control, and physiological measures such as body temperature or galvanic skin response. A benefit of classifying all variables into broad categories is that it increases the power of the meta-analysis since expanding the analysis to even more specific constructs would seriously limit the number of available studies. We made one exception and that is that we also ran tailored analyses on a subset of the intrapersonal measures that assessed *fundamental needs* (i.e., belonging, self-esteem, control, and meaningful existence). These fundamental needs measures included the typical need satisfaction measures that are especially designed for Cyberball [1,21,22] and conceptually related measures such as the Rosenberg Self-Esteem Scale. The reason why we did focus on this specific subset of intrapersonal variables is that the evidence supporting Williams' temporal model is to~~by~~ a large extent~~extend~~ based on studies using these specific dependent variables. In other words, these fundamental needs measures are particularly important for testing Williams's [11] prediction concerning moderation of ostracism effects over time.

Hypotheses

Following our preregistered report on OSF, we divided the hypotheses into two primary hypotheses and several secondary hypotheses. The two primary hypotheses were: is there an ordinal decrease of the ostracism effect across time of measurement? (Hypothesis 1)~~),~~ and is there an ordinal difference in the interaction effect across time of measurement (Hypothesis 2)? Secondary hypotheses regarded moderation of the ostracism effect by structural aspects of the studies, sampling aspects of the studies, and different types of dependent measures used. These hypotheses will be answered with random and mixed-effects meta-analytic models applied to all 120 studies that we were able to collate.

Method

Study inclusion criteria

First, we only considered Cyberball experiments that contained a factor that manipulated the number of virtual ball tosses obtained by the participants. For this ostracism factor we only considered the condition in which participants were ostracized by all other participants and the condition in which participants were equally included by all other players. Second, we only considered experiments that incorporated a between-subjects design with random assignment. Within-subject designs were excluded, because this would require the correlations between measures in primary studies and such correlations are often not reliably reported in the papers. Moreover, most within-subjects designs regard high-dimensional neurophysiological measurements such as fMRI that are beyond the scope of this meta-analysis [15,16]. Third, we checked whether the experiments contained other factors besides the ostracism factor. If the experiment contained more than two additional factors we collapsed effects sizes across the factor that authors expressed least interest in. Moreover, continuous variables that were dichotomized into factorial levels were also collapsed due to the many problems dichotomization can cause (e.g., underestimation of effect size, spurious effects [23,24]; four cases).~~.~~ Fourth, for the dependent measures the criterion was that they were (expected to be) affected by the ostracism manipulation. We considered the measures that immediately followed the manipulation (first measure) and the measure at the end of the study (last measure), while excluding manipulation checks in this assessment.

Reasons for these inclusion criteria are threefold: (1) Most Cyberball experiments take place in such a format, making it an encompassing criterion for the purposes of this meta-analysis. (2) The choice to limit the meta-analysis to between-subject designs rendered computational aspects more feasible based on reported statistics in papers. (3) The criteria maximize experimental rigor as they minimize the need for subjective quality assessment of the primary studies. Indeed, clear inclusion criteria decrease variability due to design characteristics, which increases power for moderator analyses [25].

Literature search

To have a comprehensive meta-analysis of Cyberball studies, we used seven search strategies in the period of November 2012 through April 2013. These search strategies included database searches, a call for data, cross-reference with Kip Williams's online list of Cyberball studies, Google Scholar alerts, citation records, Society for Personality and Social Psychology (SPSP) conference abstracts, and personal communications.

The databases searched included Web of Knowledge, PubMed, ScienceDirect, and Worldcat using all sources from the Tilburg University library. The first three cover only published articles, whereas Worldcat also covers books and dissertations as well as the PsycINFO database. All these databases were searched with the keywords *cyberball*, *ball-tossing* and *ball AND ostraci**. Web of Knowledge was the first database searched. For this database, an additional search term (i.e., *ball AND exclu**) was used, but this additional search term yielded zero relevant hits that were not a result of the other searches and was dropped. Across all these searches, results included 1927 potentially relevant studies of which a total of 109 were deemed relevant and saved for coding. Within Web of Knowledge, we looked through all citation records of the seminal papers by Williams et al. [1]; Williams and Jarvis [26]. These papers were cited 332 times (as of 5th of November, 2012), of which 43 papers were saved for coding. The entire literature search provided 2259 potentially relevant studies (including possible duplicates across searches), of which 152 were selected to be included in the coding.

The call for data was put on the list servers or forums of ~~Society for Personality and Social Psychology (~~SPSP~~,),~~ European Association of Social Psychology (EASP), and Social Psychology Network (SPN; all on 3rd of December, 2012). This resulted in 9 replies, yielding 3 useful studies.

Kip Williams keeps a list of Cyberball studies on his website. This list was used to check for extra articles that did not turn up in the initial searches on November 15th, 2012.[2] The list included 93 papers, of which 9 papers were included to be coded.

The final searches included Google Scholar alerts, SPSP conference abstracts, and personal communication. The Google Scholar alerts were used to keep up to date with new literature. These alerts notify a user when new search results for a search term occur and were used for *cyberball* and *ball-tossing*. This yielded 85 search results of which 25 were saved for coding. SPSP conference abstracts from 2006 through 2013 were searched for Cyberball studies. This led to personal communications with the authors of the conference abstracts, leading to additional studies. Pooled, the personal communication and the conference abstracts yielded 21 potentially relevant studies, of which 20 were saved for coding. The seminal paper by Williams et al. [1] was added separately.

In sum, the literature search spanned 2468 potentially relevant studies, resulting in 205 that were saved for coding. During coding, papers were assessed to fit the inclusion criteria. Of the 205 papers, 107 papers were excluded for a variety of reasons. See also Fig. 2. Several involved the use of a within-subjects design (52 papers). Some papers could not be accessed (5 papers) or could not be included because we did not receive the required data on request (7 papers). Some were excluded for other reasons (43 papers), such as not involving new data (e.g., a dissertation study that was later published). All included papers were published between 2000 (after the introduction of Cyberball) and April 2013. This resulted in a final, fully coded sample of 98 papers containing 120 studies, with mean sample size 98.9 and median sample size 74.[3] There were a total of 11,869 Cyberball participants.

Formatted: Not Superscript/ Subscript

Fig. 2. PRISMA flowchart of the current meta-analysis.

Coding procedure

The first author coded all the studies and conducted all the analyses. The second author double-checked the coding of all 52 studies that entailed a full two-by-two design. The third author double-checked and reran the R code of all analyses. Finally, an extensive account of all coding decisions is publicly available via Open Science Framework on a paper-by-paper basis (see Footnote 2 for the direct link, Supplement S1 also contains the data).

We first coded the structural aspects and sample aspects of all papers. The structural aspects of Cyberball that we coded were (1) number of players depicted in Cyberball, (2) total number of ball tosses used throughout the game, (3) total duration of the game in seconds. The sample aspects that we coded were (1) percentage of male participants, (2) average age of participants, and (3) country of origin.

We then coded the dependent variables that were relevant for the current meta-analysis by retrieving the means and standard deviations of the first and the last relevant measure of all papers. Importantly, to estimate the duration between the first and last measure we counted the number of questions that were assessed between the two measures. Specifically, following a longstanding practice in the freshman testing program of the University of Amsterdam [27] we estimated that participants would need 6 seconds on average to complete one question. Moreover, we included additional time if this was explicitly reported in the method section of the manuscript or when a measure would clearly deviate from 6 seconds to complete (e.g., tasks that measure endurance such as a grip strength task).

Both first and last measures were subsequently coded in the following general terms: (1) interpersonal, (2) intrapersonal, (3) fundamental needs, (4) model correspondence. Interpersonal measures were defined as measuring constructs that relate to (the self and) others (e.g., *how angry do you feel towards person X?*, donations to charity).~~, etc.).~~ Intrapersonal measures were defined as measuring constructs that relate only to the self (e.g., *how angry do you feel?*, physiological measures).~~, etc.).~~ Fundamental needs measures were those that measured self-esteem, belonging, control, meaningful existence, or a composite of these. Note that the fundamental needs are a refinement of the intrapersonal measures and that intrapersonal measures thus include the fundamental need measures. The model correspondence variable coded whether the first- and last measure fit the definition William's ostracism model that a variable can indeed be classified as an immediate measure (i.e., during the game) and delayed measure (i.e., after the game/now), respectively.

The consequence of including many different kinds of dependent variables is that some measures are expected to increase as a function of ostracism (e.g.,~~.~~ need threat) and others are expected to decrease (e.g., need satisfaction). To counteract computational problems (i.e., cancellation of effects) being caused by this bidirectionality of ostracism effects, we coded the direction of the ostracism effect for each specific measure, such that negative effect sizes depict negative psychological effects.

A similar argument can also be made about including multiple moderator variables in the analysis of interaction effects. In the 52 studies that included a moderator variable we thus needed to account for the expected direction of every moderator. If we had not done this, the interaction effects could cancel out, thereby leading to ambivalent results. To explain this, we present in Table 1 hypothetical data for the four different study designs that are possible when crossing direction of the effect and direction of the moderation. The relevant effect sizes should be corrected to attain comparable effect sizes across studies. Effect sizes for the simple ostracism effect (column wise) were corrected only for the type of measure. For instance, for panels (a) (involving, e.g., need threat) and (c) (involving, e.g., need satisfaction), the corrections entailed a multiplication with -1 or +1, respectively. Simple moderator effects (row wise comparisons) are interesting for understanding the effect of the moderator under either ostracism or inclusion. These simple moderator effects were corrected for both the type of measure *and* the expected moderation (i.e., exacerbation, -1, or minimization, +1). For example in panel (c), the 5 and 8 on the right are used to compute the *standard ostracism effect* (as in [1]), whereas the 3 and 8 in the left column represent an ostracism effect that is thought to be exacerbated. For example, in a given ostracism study with a two-by-two design, adolescents are expected to show stronger ostracism effects, compared to young adults [19]. The 5 and 8 would subsequently represent the scores for the young adults, whereas the 3 and 8 would represent the scores for the young adolescents. In panel (d) we depict a study in which the *moderated* column is thought to lead to a minimal ostracism effect, as could be expected when Cyberball is played with members of a despised out-group [28]. The margins (greyed out) denote the simple effects, which are after correction comparable across all panels (a) through (d), indicating that this correction did what we intended it to.

Table 1. Hypothetical data example of coding correction.

(a) Negative moderator, negative measure

		Moderated	Not-moderated/control	Raw	Correct
Ostracism factor	Ostracism	13	11	2	2
	Inclusion	8	8	0	0
	Raw	5	3		
	Correct	-5	-3		

(b) Positive moderator, negative measure

		Moderated	Not-moderated/control	Raw	Correct
Ostracism factor	Ostracism	9	11	-2	2
	Inclusion	8	8	0	0
	Raw	1	3		
	Correct	-1	-3		

(c) Negative moderator, positive measure

		Moderated	Not-moderated/control	Raw	Correct
Ostracism factor	Ostracism	3	5	-2	2
	Inclusion	8	8	0	0
	Raw	-5	-3		
	Correct	-5	-3		

(d) Positive moderator, positive measure

		Moderated	Not-moderated/control	Raw	Correct
Ostracism factor	Ostracism	7	5	2	2
	Inclusion	8	8	0	0
	Raw	-1	-3		
	Correct	-1	-3		

Raw denotes the simple effect in the hypothetical data before correction whereas correct denotes the simple effect after correction. Column wise effects are multiplied by the type of measure only, whereas row~~column~~ wise effects are multiplied by both the type of moderator and type of measure.

Finally, relevant information that was missing in the papers was requested from the authors via e-mail. In case of non-response, we sent three follow-up e-mails. All this communication was documented and can be found on the OSF page for this project. In case of non-response or non-willingness to send data, studies were either eliminated if the information was crucial (i.e., means and standard deviations of the measures per group), computed if possible (i.e., cell sizes), or assumed if deemed reasonable on the basis of additional information. For instance, when no information was given we considered the Cyberball manipulation characteristics to be similar to previous studies in the same paper or in earlier papers referred to in the paper (descriptions of all cases are described in the log file on the OSF).

Statistical analyses

For the analyses, we used version 1.9-5 of the *metafor* package [29] in the R statistical environment [30].

Effect size metric

We used Hedges's *g* version of the standardized mean differences as the effect size. Hedges's *g* corrects for the slightly biased estimate given by Cohen's *d* [31]. Standardized ~~simple~~ effects were calculated across the ostracism factor, where ~~and~~ the 52 studies with a cross-cutting variable were included as a simple effect of ostracism within the non-moderated level. Standardized interaction effect were~~was~~ calculated by taking the standardized difference between the unstandardized main effects (see the Appendix for the exact formulae used). These effects were computed~~This was done~~ for both the first and last dependent variable in each experiment. For example, in a 2 (ostracized vs. included) by 2 (moderator present vs. moderator absent) design with multiple measures, we calculated two simple ostracism effects (Hypothesis 1) and two interaction effects (Hypothesis 2). For ten studies, more factors/levels were used and a 2 by 2 was extracted~~Non-factorial studies delivered only simple effects for the first and last measure and no interactions~~.

Meta-analytic model

We used random- and mixed-effects models, because heterogeneity in the effect sizes is expected due to both the inclusion of different measures and additional unknown methodological and substantive factors. The meta-regression element in some of the analyses is the variable time as predictor of the ostracism effect. Analyses without this study-level predictor reduce to a random-effects model. We used Restricted Maximum Likelihood (REML) to estimate tau-squared (i.e., the residual variance), as recommended by Viechtbauer [32]. Note that when estimating a mixed- or random effects model, one does not estimate a single *true* effect, but rather the mean and variance of underlying effects [32].

Statistical sensitivity analyses

To test for robustness of the effects, we incorporated several statistical sensitivity analyses. We flagged possibly problematic outliers on the basis of studentized deleted residuals, Q-Q plots, and Cook’s distance values. Subsequently, we inspected the effect of these outliers on substantial results in statistical sensitivity analyses in which these outliers were excluded. Another statistical sensitivity analysis entailed fitting of the mixed-effects model with tau-squared fit at the upper bound value of the 95% confidence interval.

Funnel plot asymmetry

A funnel plot depicts each study’s effect size against its standard error [33]. Larger studies have smaller standard errors, and vice versa for smaller studies. Following from a theoretical fluctuation of the population effect size due to sampling variance, a funnel plot should be symmetrical around the estimated mean effect size. If there are no methodological or substantive reasons to expect a link between effect sizes and standard errors, funnel plot *asymmetry* can indicate publication bias (e.g., [34]). To test funnel plot asymmetry, we used

Egger's regression test [35] for mixed-effects models [36].[4] This tests whether the distribution of effect sizes is equal on both sides of the average effect, when accounting for true heterogeneity. Funnel plot asymmetry thus indicates bias in the estimated mean effect size~~,~~ and possibly publication bias.

Results

In our reporting of the effect sizes, *d* indicates a main effect and Δ*d* indicates an interaction effect. Even though we used Hedges's *g*, we maintained the notation of *d*, because *g* is only a minor correction to Cohen's *d*. Statistical sensitivity analyses are only reported if they showed different effects (all statistical sensitivity analyses can be found on OSF).

Primary analyses

The two primary hypotheses are tested in four meta-analyses, of which the study level effects are reported in Table 2. The table includes effect sizes used in the estimation of the average simple effect of ostracism on the first measure, the average simple effect on the last measure and the estimation of the average interaction effect on both the first and last measure.

Table 2. Effect sizes per study for the primary hypotheses.

First author	Year	*N*	*d* T1	(*SE*)	*d* T2	(*SE*)	Δ*d* T1	(*SE*)	Δ*d* T2	(*SE*)
Alvares	2010	74	-1.21	0.12	-0.10	0.10	-0.15	0.24	1.12	0.23
Ambrosini	2013	40	-1.69	0.13	-0.97	0.11	-	-	-	-
Aydin	2012	68	-0.95	0.13	-0.40	0.12	-1.19	0.24	0.72	0.23
Banki	2012	89	-1.87	0.07	-0.35	0.05	-	-	-	-
Bastian	2010	72	-2.75	0.11	-1.42	0.07	-	-	-	-
Bernstein	2012	24	-0.41	0.16	-	-	-	-	-	-
Bernstein	2012	25.50	-1.04	0.17	-	-	-	-	-	-
Bernstein	2010	73	-1.63	0.16	-1.63	0.16	-0.86	0.37	-1.11	0.40

First author	Year	*N*	*d* T1	(*SE*)	*d* T2	(*SE*)	Δ*d* T1	(*SE*)	Δ*d* T2	(*SE*)
Bernstein	2010	138	-2.67	0.10	-1.96	0.08	-0.53	0.22	-0.51	0.17
Bernstein	2012	67	-2.00	0.17	-0.99	0.13	-1.07	0.45	-0.80	0.30
Bernstein	2012	27	-1.39	0.17	-	-	-	-	-	-
Boyes	2009	89	-0.43	0.05	-0.80	0.05	-	-	-	-
Boyes	2009	87	-0.20	0.05	-0.84	0.05	-	-	-	-
Brochu	-	35	-2.51	0.20	-0.48	0.11	-	-	-	-
Brown	2009	52	-0.64	0.08	-	-	-	-	-	-
Carter	2008	143	-0.28	0.06	0.20	0.06	0.34	0.11	0.17	0.11
Carter-Sowell	2008	65	-2.86	0.12	-1.48	0.08	-	-	-	-
Carter-Sowell	2010	74	-1.60	0.14	-1.49	0.13	-1.23	0.33	-1.15	0.34
Carter-Sowell	2010	70.67	-2.09	0.17	-0.56	0.11	-0.65	0.39	-0.63	0.24
Chen	2012	60	-1.04	0.14	-	-	-1.35	0.27	-	-
Chen	2012	83	-1.32	0.11	-	-	-1.32	0.21	-	-
Chernyak	2010	76	-1.52	0.10	0.15	0.08	-	-	-	-
Chow	2008	75	-1.20	0.06	-1.31	0.06	-	-	-	-
Chrisp	2012	77	-0.70	0.06	-0.15	0.05	-	-	-	-
Coyne	2011	40	-0.56	0.10	-	-	-	-	-	-
De Waal-Andrews	2012	136	-3.55	0.16	-2.55	0.11	-1.29	0.24	-0.87	0.18
De Waal-Andrews	2012	112	-4.21	0.22	-2.17	0.11	-1.56	0.31	-1.20	0.18
DeBono	-	57	-1.07	0.15	-0.05	0.13	-1.55	0.29	-0.48	0.27
DeBono	-	81	-1.07	0.11	-0.10	0.09	-0.33	0.21	0.24	0.19
DeBono	-	83	-0.13	0.09	-	-	-0.75	0.19	-	-
Dietrich	2010	75	1.43	0.07	-	-	-	-	-	-
Duclos	2012	59	-0.63	0.07	-	-	-	-	-	-
Eisenberger	2006	48	-0.15	0.08	-1.24	0.10	-	-	-	-
Fayant	-	60	-2.04	0.20	-1.12	0.15	0.22	0.38	-0.44	0.28
Floor	2007	88	-1.92	0.13	-0.73	0.09	-0.21	0.28	-0.59	0.19
Gallardo-Pujol	2012	57	-1.18	0.16	-0.52	0.15	-1.17	0.31	0.11	0.29
Gan	2012	72	-0.54	0.03	-0.07	0.03	-0.62	0.06	0.02	0.06

First author	Year	*N*	*d* T1	(*SE*)	*d* T2	(*SE*)	Δ*d* T1	(*SE*)	Δ*d* T2	(*SE*)
Garczynski	2013	83	-1.51	0.19	0.39	0.15	-1.29	0.33	-0.01	0.29
Geniole	2011	74	0.19	0.06	-0.11	0.06	-	-	-	-
Gerber	-	38	-2.09	0.16	-	-	-	-	-	-
Gerber	-	89	-3.38	0.21	-	-	-	-	-	-
Gonsalkorale	2007	97	-1.31	0.14	0.26	0.12	0.49	0.30	1.31	0.25
Goodwin	2010	300	-1.81	0.04	-0.94	0.03	0.20	0.08	-0.43	0.07
Goodwin	2010	314	0.13	0.02	-0.09	0.02	0.35	0.06	-0.10	0.06
Greitemeyer	2012	56	-0.48	0.07	-0.23	0.07	-	-	-	-
Gruijters	-	113	-0.26	0.06	-1.07	0.07	-	-	-	-
Hackenbracht	2013	51	-1.92	0.11	-0.18	0.08	-	-	-	-
Hawes	2012	55	-2.16	0.23	0.69	0.15	0.00	0.38	-1.05	0.28
Hellmann	-	76	-1.21	0.12	0.19	0.10	-1.40	0.22	0.74	0.21
Hess	2010	162	-2.34	0.04	-0.87	0.03	-	-	-	-
Hess	2011	38	-0.64	0.11	-	-	-	-	-	-
Horn	-	68	-0.77	0.12	-0.99	0.13	-0.99	0.23	1.49	0.24
IJzerman	2012	86	-1.67	0.12	-	-	-1.07	0.22	-	-
Jamieson	2010	33	-1.56	0.15	-1.06	0.13	-	-	-	-
Jamieson	2010	68	-1.94	0.09	-1.47	0.07	-	-	-	-
Johnson	2010	104	-0.73	0.04	-0.79	0.04	-	-	-	-
Kassner	-	85	-1.72	0.13	-1.02	0.11	-0.87	0.31	-0.30	0.21
Kassner	2012	49	-2.11	0.12	-1.78	0.11	-	-	-	-
Kerr	2008	250	-1.66	0.02	-0.05	0.02	-	-	-	-
Kesting	2013	76	-0.28	0.05	-0.79	0.06	-	-	-	-
Knowles	2010	62	-0.38	0.12	-	-	-0.99	0.25	-	-
Knowles	2012	60	-0.60	0.07	-	-	-	-	-	-
Krijnen	2008	144	-4.74	0.11	-0.18	0.03	-	-	-	-
Krill	2008	119	-2.11	0.05	-0.57	0.03	-	-	-	-
Lakin	2008	36	-1.53	0.14	-0.51	0.11	-	-	-	-
Lau	2009	56	-2.50	0.23	-1.09	0.15	-0.06	0.58	1.36	0.46

First author	Year	*N*	*d* T1	(*SE*)	*d* T2	(*SE*)	Δ*d* T1	(*SE*)	Δ*d* T2	(*SE*)
Lustenberger	2010	71	-0.83	0.06	0.04	0.06	-	-	-	-
Lustenberger	2010	156	-0.70	0.03	-	-	-	-	-	-
MacDonald	2008	63	-0.15	0.06	-	-	-	-	-	-
McDonald	2012	270	-0.06	0.02	-2.40	0.03	-	-	-	-
Nordgren	2011	71	-0.74	0.06	-	-	-	-	-	-
Nordgren	2011	74	-0.80	0.06	-	-	-	-	-	-
Nordgren	2011	46	-2.24	0.14	-	-	-	-	-	-
Nordgren	2011	44.67	-0.55	0.09	-0.75	0.09	-	-	-	-
Nordgren	2011	58.67	-0.65	0.07	-	-	-	-	-	-
Oberleitner	2012	88	-2.36	0.08	0.42	0.05	-	-	-	-
O’Brien	2012	125	-0.58	0.03	-0.69	0.03	-	-	-	-
Peterson	2011	40	-0.89	0.11	-0.91	0.11	-	-	-	-
Pharo	2011	74	-1.33	0.13	-0.58	0.11	-1.01	0.30	-0.84	0.23
Plaisier	2012	149	-0.36	0.05	0.23	0.05	-0.40	0.11	-0.56	0.11
Ramirez	2009	121	-2.26	0.05	-1.02	0.04	-	-	-	-
Ren	2012	53	-2.18	0.12	-0.17	0.07	-	-	-	-
Renneberg	2011	60	-1.46	0.16	-1.30	0.15	0.47	0.29	0.51	0.29
Riva	2011	100	-2.10	0.13	-1.09	0.09	-	-	-	-
Ruggieri	-	91	-0.39	0.04	-0.57	0.05	-	-	-	-
Ruggieri	-	74	-0.06	0.13	-0.23	0.13	-0.31	0.24	-0.68	0.23
Sacco	2011	51	-2.40	0.13	-1.45	0.10	-	-	-	-
Sacco	2011	21	-2.28	0.29	-1.46	0.22	-	-	-	-
Sacco	2011	38	-1.74	0.14	-1.04	0.11	-	-	-	-
Salvy	2010	59	-1.45	0.08	-1.43	0.08	-	-	-	-
Salvy	2009	103	-1.48	0.05	-1.31	0.05	-	-	-	-
Schaafsma	2012	720	-1.42	0.02	-0.49	0.02	0.09	0.03	0.33	0.03
Segovia	2012	56	0.14	0.13	-	-	-1.89	0.32	-	-
Staebler	2011	68	-0.79	0.12	-0.05	0.12	0.50	0.23	0.42	0.23
Stillman	2009	121	-0.74	0.15	-1.13	0.16	0.57	0.22	-1.19	0.24

First author	Year	*N*	*d* T1	(*SE*)	*d* T2	(*SE*)	Δ*d* T1	(*SE*)	Δ*d* T2	(*SE*)
Stock	2011	155	-2.00	0.04	-0.13	0.03	-	-	-	-
Van Beest	2011	87	-0.94	0.10	-0.58	0.09	-0.40	0.24	-0.44	0.19
Van Beest	2011	183	-2.64	0.13	-0.50	0.07	-0.76	0.22	-0.11	0.13
Van Beest	2006	135	-1.29	0.07	-0.65	0.06	-0.10	0.14	-0.13	0.12
Van Beest	2006	111.33	-2.11	0.11	0.09	0.07	-0.09	0.22	-0.19	0.14
Van Beest	2012	125	-2.68	0.11	-1.24	0.07	0.06	0.35	-0.23	0.15
Van Beest	2012	85	-3.10	0.20	0.05	0.09	-0.28	0.44	0.07	0.18
Van Beest	2013	49	-3.97	0.24	-1.32	0.10	-	-	-	-
Van Beest	2013	91	-3.17	0.20	-0.48	0.09	0.75	0.56	0.53	0.18
Van Dijk	-	51	-1.50	0.10	-0.04	0.08	-	-	-	-
Webb	-	170	-0.91	0.05	-0.38	0.05	0.03	0.10	0.04	0.09
Weik	2010	65	0.16	0.12	-0.22	0.12	-0.43	0.24	0.66	0.24
Wesselmann	2009	82	-0.71	0.10	-2.03	0.14	-1.30	0.24	-0.20	0.28
Wesselmann	2012	91	-1.46	0.06	-	-	-	-	-	-
Williams	2002	390	-0.39	0.01	-2.35	0.02	-	-	-	-
Williams	2000	732	-0.79	0.01	-1.44	0.01	-	-	-	-
Williams	2000	111	-0.26	0.06	-1.01	0.07	-0.20	0.15	-0.98	0.15
Wirth	2009	159.33	-2.29	0.08	-0.76	0.05	0.05	0.17	0.46	0.11
Wirth	2010	76	-0.96	0.06	-1.64	0.07	-	-	-	-
Zadro	2004	62	-1.63	0.16	-0.19	0.12	-0.11	0.32	-1.12	0.28
Zadro	2004	77	-1.75	0.14	-0.33	0.10	-0.29	0.28	-0.70	0.21
Zadro	2006	56	-3.70	0.19	-0.87	0.08	-	-	-	-
Zhong	2008	52	-0.72	0.15	-	-	-	-	-	-
Zoller	2010	57	-0.24	0.07	-0.09	0.07	-	-	-	-
Zwolinski	2012	56	-2.01	0.11	-0.28	0.07	-	-	-	-

d T1 refers to ostracism effect on first measure; *d* T2 refers to ostracism effect on last measure; Δ*d* represent interactions. Multiple rows for the same first author and year is possible due to multiple studies across papers. Non-integer *N*s arise from division of full sample N for included conditions, appropriate due to random assignment (e.g., two conditions out of 3, when sample is 56: (56 / 3) × 2 = 37.333).~~.~~ Supplement S2 gives the full reference list of the papers in this table.

Simple ostracism effect (Hypothesis 1)

In a random-effects model on the main effect of ostracism (k = 120), residual heterogeneity was significant, Q (119) = 1395, p < .001, I^2 = 92.99% and estimated at τ^2 = 0.90, 95% CI [0.70, 1.24]. The heterogeneity measure τ^2 includes both the estimated proportion of explained variance at the study level and unexplained variance in the distribution of underlying effect sizes (i.e., τ_{res}^2). The analysis yielded an estimated average effect of d = -1.36, p < .001, 95% CI [-1.54, -1.18]. A random-effects version of the Egger's test [36] indicated funnel plot asymmetry, Z = -6.14, p < .001. Due to the size of the average effect,~~ and~~ hence large power to acquire significant outcomes in primary studies, we do not suspect publication bias to explain this asymmetry. In other words, immediately after being ostracized, the average ostracism effect is estimated at -1.36 standard deviation units, which entails a large effect [37].

Next, we fitted a mixed-effects regression model for the ostracism effect on the last measure (k = 95), including estimated time in seconds since completing the Cyberball game as predictor. Residual heterogeneity was significant, Q_E (93) = 803, p < .001 and estimated at τ_{res}^2 = 0.38, 95% CI [0.27, 0.54]. The intercept was estimated at $d_{intercept}$ = -0.76, p < .001, 95% CI [-0.91, -0.61]. Moreover, the estimated time in seconds between exclusion in Cyberball and the moment at which the last measure was taken failed to moderate the average effect, b = 0.0069~~0001~~, p = .187, 95% CI [-0.0034~~0001~~, 0.0172~~0003~~]. However, we have to take into consideration the low power of the moderation analyses due to the large (residual) heterogeneity in effect sizes [25]. A regression test for mixed-effects model with moderator (i.e., including both the time and *SE* as predictor) showed no funnel plot asymmetry, Z = -0.72, p = .474. In short, long after ostracism has occurred (M_{time} = 4.85 minutes~~291.2 seconds~~), ostracized participants on average scored around -0.73 standard deviation units lower when compared with included participants, an effect that does not appear to be moderated further by time passed since the ostracism occurrence.

Thus, results show a clear effect of ostracism on both the first and last measures, of which the latter is *not* predicted by our operationalization of time. The ostracism effect over time can also be inspected via confidence intervals. Comparing the 95% confidence intervals for the average ostracism effect on the first measure (i.e., [-1.54, -1.18]) and on the last measure (i.e., [-0.86, -0.59]) showed no overlap. Although the difference in average effect sizes between first and last measure cannot be formally tested (because of a lack of information on the correlation between measures in the primary studies), the mean difference is sizeable and CIs confirms our prediction that the average ostracism effect is smaller for the last measure. In fact, given the expected positive correlation between effects for first and last measures, the comparison of CIs is likely to be conservative [38]. Additionally, we noted that estimated residual heterogeneity was larger on the first- than on the last measure. We conclude that the average ostracism effects decreases from the first- to last measures~~,~~ and that study-level effects are more similar on the last measure.

Moderation of ostracism (Hypothesis 2)

To test moderation of the ostracism effect, we selected the factorial experiments that manipulated ostracism and another independent variable in between-subjects designs. A random-effects model on the interaction effect (Δd) on the first measure (k = 52) showed heterogeneity in underlying effects, Q (51) = 103.24, p < .001, I^2 = 50.60% and an estimated τ^2 = 0.19, 95% CI [0.07, 0.41]. The average interaction effect equaled Δd = -0.46, p < .001, 95% CI [-0.64, -0.28], indicating a change in the ostracism effect due to the moderator level and vice versa (i.e., moderation of the ostracism effect). There was indication of funnel plot asymmetry in this analysis, Z = -2.43, p = .015. Thus, the data indicate that, across the board, the ostracism effect *can* be moderated on the first measure following the ostracism sequence, but it is possible that publication bias may have affected the interaction estimates.

On the last measure (k = 46), the mixed-effects model (with estimated time as predictor) for the interaction effect again showed residual heterogeneity, Q_E (44) = 100.82, p < .001 and estimated τ_{res}^2 = 0.21, 95% CI [0.10, 0.55]. The intercept of the interaction effect was estimated at $\Delta d_{intercept}$ = -0.20, p = .052, 95% CI [-0.402, 0.002] and no significant moderation of time was found, b = 0.011~~0002~~, p = .159, 95% CI [-0.0043~~0001~~, 0.0264~~0004~~]. The regression test with the time and SE as predictors showed no funnel plot asymmetry, Z = -0.68, p = .495. These results indicate that moderation of the average ostracism effect is *not* found at a later time point in the included studies~~,~~ and time itself does not moderate the computed interaction effects. However, statistical sensitivity analyses showed that this interaction *was* significant when we removed three outliers based on studentized residuals, $\Delta d_{intercept}$ = -0.32, p = .029, 95% CI [-0.60, -0.03], whereas the regression coefficient time continued to be non-significant, b = 0.0002, p = .207, 95% CI [-0.0001, 0.0006]. On the last measure, this indicates that the non-significant interaction effect is sensitive to outliers in the data.

To see whether the interaction effects changed from the first to the last measure, we again compared confidence intervals. On the first measure, the 95% CI was [-0.64, -0.28] whereas for the last measure, the 95% CI was [-0.32, 0.05]. Considering the overlap of these CIs, one needs to be careful to interpret this as a reduction in the moderation across the measures examined. It is clear, however, that the average effect size of the interaction does not increase from first to last measure.

Secondary analyses

In addition to the simple effects over all studies, we analyzed subsets of studies that differ in type of dependent measure to study robustness of the effects. We also inspected whether sample composition, scale composition, and Cyberball specifics could predict the estimated effect size. Finally, we selected a homogeneous subset of studies to come to grips with the relatively large heterogeneity of simple main effects found for the primary hypotheses.

Measures

To inspect the robustness of the estimates of the first and last measure, we studied simple effects across several subsets of measures. These subsets encompassed interpersonal measures (i.e., measures that relate to others or the self in the context of others), intrapersonal measures (i.e., measures that relate only to the self), fundamental needs (single- and composite needs), and measures that were coded by the first two authors as fitting the description of being immediate or delayed (i.e., questions related to during- or after the game, respectively; shown in Fig. 3~~2~~ as *model*). We ran the analyses for the different measures for the two time points separately (i.e., first and last measure).

Fig. 3~~2~~. Dotplots of the average estimated simple effects with 95% confidence intervals. T1 represents first measure~~,~~ and T2 represents last measure. These effects are across the same subset. Traditional ostracism effect refers to the between-subjects effect of being ostracized with *no* moderator present, whereas moderated ostracism effect refers to being ostracized *with* a moderator present. Vice versa, moderator effect within ostracism/inclusion level refers to the between-subjects effect of the moderator factor, within the ostracized/inclusion conditions. The subset labeled “All” contains~~ =~~ all measures. The subset labeled “~~; ~~Fundamental” contains~~ =~~ only fundamental need measures. The subset labeled “~~; ~~Intrapersonal” contains~~ =~~ all intrapersonal measures. The subset labeled

“Interpersonal” contains~~; interpersonal =~~ all interpersonal measures. The subset labeled “Model” contains those where~~; model =~~ first measures is immediate and last measure is delayed. See~~For lists of studies in each subset, see~~ Supplement S4~~S3~~.

The different panels in Fig. 3~~2~~ show the results for the different simple effects per subset and overall; Table 3 summarizes the estimated interaction effects. A comparison of the results within each panel shows whether the overall results are robust and representative of all subsets, or whether there are nuances per type of measure. The main differences are notable in panels (1), (2),~~)~~ and (5). The first and second panels indicate that the effect of ostracism is weaker for interpersonal measures, compared to all intrapersonal measures (including fundamental needs). This indicates that in a similar factorial design, interpersonal measures show weaker effects than intrapersonal measures. Panel 5 indicates that the moderation of interpersonal measures is stronger compared to the other subsets. This suggests that interpersonal measures are more subject to moderation, whereas the effects of ostracism on interpersonal measures are smaller initially. Additionally, for the specific subset of fundamental needs, we noted that the point estimated interactions (Table 3) follow the pattern predicted by the need-threat model [11]: ~~i.e.,~~ the first measures are moderated less strongly than the last measures.[5~~4~~]

Table 3. Interaction effect per subset.

		k	Estimate	(*SE*)	*Z*-value	*p*-value	95% CI	95% CI
							Lowerbound	Upperbound
Overall	T1	52	-0.46	0.09	-5.08	< .001	-0.64	-0.28
	T2	46	-0.19	0.11	-1.82	.069	-0.40	0.02
Fundamental	T1	30	-0.39	0.12	-3.42	< .001	-0.62	-0.17
	T2	17	-0.77	0.25	-3.05	.002	-1.27	-0.28

Intrapersonal	T1	42	-0.31	0.09	-3.38	< .001	-0.49	-0.13
	T2	39	-0.21	0.11	-1.87	.062	-0.44	0.01
Interpersonal	T1	10	-1.03	0.18	-5.69	<.0001	-1.38	-0.67
	T1$_{\text{listwise}}$	6	-0.36	0.22	-1.63	.104	-0.79	0.07
	T2	6	0.63	0.62	1.02	.309	-0.58	1.84
Model	T1	36	-0.29	0.10	-2.99	.003	-0.48	-0.10
	T2	23	0.01	0.17	0.08	.938	-0.31	0.34

The subset labeled “All” contains all measures. The subset labeled “Fundamental” contains only fundamental need measures. The subset labeled “Intrapersonal” contains all intrapersonal measures. The subset labeled “Interpersonal” contains all interpersonal measures. The subset labeled “Model” contains those where first measures is immediate and last measure is delayed. See Supplement S4. Listwise deletion ensures that estimates are made on full rows in the data. Listwise deletion was applied in all the subsets, which only altered results for interpersonal measures.~~Overall estimates are based on all data, where the rest form subsets. Model indicates that the first measure was indeed reflexive and the last measure reflective. Listwise deletion for equal *ks* across time points within a subset yielded highly similar results, except for interpersonal measures, which is depicted in the row labeled T1~~$_{\text{listwise}}$~~.~~

Composition

To inspect for structural and sampling effects of the studies, we~~We~~ ran mixed-effect~~effects~~ models on the 120 ostracism ~~effect (as in Hypothesis 1) inspecting for composition~~ effects, on both the first and the last measure. Due to listwise deletion, only 45 of 120 effect sizes remained on the first measure and 41 of 95 effect sizes for the last measure. The predictors in the mixed effects model were (1) country (US, other Western country, Asian, other), (2) proportion of males in the study, (3) mean age of the sample, (4) number of players in the game, (5) length of the game (≤ 5min, 5-10 min or > 10 min), (6) the number of throws in the game and (7) type of needs scale referenced (by assigning unique values for every unique reference).

On the first measure, this model (k = 45) showed clear residual heterogeneity after controlling for these structural- and sampling aspects of the studies, Q_E (33~~32~~) = 449.52, p < .001, estimated τ_{res}^2 = 0.90, 95% CI [0.54, 1.59], but no overall moderation, Q_M (11) = 10.75, p = .465. The different types of need scales [11,21,22] did not significantly moderate effect sizes, showing psychometric convergence among the three scales. Inspecting the predictors individually also showed no indication for moderation (ps > .137; see Table 4).

Table 4. Meta regression coefficients for composition effects (first measure; k = 45).

	Estimate	(*SE*)	*Z*-value	*p*-value	95% CI Lowerbound	95% CI Upperbound
Intercept	-2.14	3.27	-1.89	0.058	-4.35	0.07
Structural						
Nr. of players	-0.22	1.05	-0.21	0.837	-2.28	1.85
Nr. of throws	0.03	0.02	1.49	0.137	-0.01	0.07
Ostracism <5 min	-	-	-	-	-	-
Ostracism 5-10 min	0.75	0.81	0.92	0.358	-0.84	2.34
Need scale = Williams (2000)	-	-	-	-	-	-
Need scale = Zadro et al. (2004)	-0.36	0.41	-0.88	0.381	-1.16	0.45
Need scale = Van Beest & Williams (2006)	0.07	0.54	0.13	0.894	-0.98	1.12
Need scale = Williams Zadro	-0.03	0.62	-0.04	0.965	-1.25	1.19
Need scale = Gonsalkorale &	0.68	0.82	0.82	0.414	-0.94	2.30

Williams (2007)						
Sampling						
Country = US	-	-	-	-	-	-
Country = Western	-0.42	0.36	-1.15	0.249	-1.13	0.29
Country = Asian	-0.30	1.13	-0.26	0.793	-2.51	1.92
Proportion male	1.54	1.09	1.42	0.156	-0.59	3.68
Mean age	-0.05	0.05	-0.97	0.332	-0.16	0.05

This can be interpreted as a standard regression formula. Empty rows represent reference categories.

On the last measure (k = 41; Table 5), no overall moderation was found, Q_M (11~~12~~) = 6.00, p = .873, but heterogeneity did occur, Q_E (29) = 214.69, p < .0001. The ~~the~~ number of players in the game ~~did~~ significantly predicted~~predict~~ the effects, b = 1.55, p = .047, 95% CI [0.2; 3.07], which would be interpreted as four players eliciting smaller ostracism effects, when compared to three players. The significance of this individual predictor should be interpreted carefully, as the omnibus moderation test showed no systematic decrease in heterogeneity. Overall, we found no strong evidence for moderation due to study or sample composition.[6~~5~~]

Table 5. Meta-regression coefficients for composition effects (last measure; k = 41).

	Estimate	(*SE*)	*Z*-value	*p*-value	95% CI Lowerbound	95% CI Upperbound
Intercept	-1.12	0.92	-1.21	0.227	-2.95	-0.70
Structural						
Nr. of players	1.55	0.78	1.98	0.047	0.02	3.07
Nr. of throws	0.01	0.02	0.59	0.556	-0.02	0.04
Ostracism <5 min	-	-	-	-	-	-

Ostracism 5-10 min	0.38	0.62	0.61	0.539	-0.83	1.59
Need scale = Williams (2000)	-	-	-	-	-	-
Need scale = Zadro et al. (2004)	-0.14	0.32	-0.44	0.658	-0.77	0.49
Need scale = Van Beest & Williams (2006)	-0.21	0.41	-0.51	0.613	-1.02	0.60
Need scale = Williams Zadro	-0.12	0.53	-0.22	0.826	-1.16	0.92
Need scale = Gonsalkorale & Williams (2007)	-0.07	0.65	-0.10	0.916	-1.33	1.20
Sampling						
Country = US	-	-	-	-	-	-
Country = Western	0.26	0.30	0.87	0.387	-0.33	0.86
Country = Asian	0.85	0.84	1.01	0.313	-0.80	2.49
Proportion male	0.29	0.83	0.35	0.730	-1.34	1.91
Mean age	-0.01	0.04	-0.25	0.806	-0.10	0.08

This can be interpreted as a standard regression formula. Empty rows represent reference categories.

Homogeneity?

The analysis of the simple ostracism effect on the first measure showed that differences of underlying effects made up 93% of the variability in study outcomes. We performed an additional secondary analysis in a more homogenous subset of studies to better understand this heterogeneity. This subset only included typical Cyberball studies that involved three players in the game, 30 throws, and lasted less than five minutes. In addition, the homogeneous subset of typical Cyberball studies only involved measures of immediate fundamental needs (single or composite). Performing a meta-analysis on this homogeneous subset of 19 studies showed an I^2 value of 83%, indicating that 83% of the total variability can be attributed to heterogeneity in the effect sizes. We noted that the mean simple ostracism effect in these 19 studies was relatively strong and estimated at d = -2.05, 95% CI [-2.44, -

1.65]. In other words, given that the heterogeneity remains large even in a homogeneous subset, suggests that the heterogeneity found in the overall analyses does not appear to be an artifact from the inclusion of different measures and the use of alternative Cyberball setups.

Discussion

In this meta-analysis of Cyberball studies we estimated the average ostracism effect of the first and last dependent variable used in 120 Cyberball experiments. The primary hypotheses were (a) that the ostracism effect size would decrease from first to last measure and (b) that first measures would be less affected by cross-cutting variables than last measures. The secondary hypotheses tested whether the above generalizes across structural variables of the game, sample characteristics, or type of dependent variable used.

The results confirmed the hypothesis that the ostracism effect decreased from the first (d = -1.36) to the last measure (d = -.76), although this decline was not predicted by our estimation of duration between first and last measure. The results did not fully confirm the hypothesis that last measures are more strongly moderated than first measures. That is, our analysis of the experiments that included an experimentally controlled cross-cutting variable revealed that cross-cutting variables moderated both the first and last measure. In fact, visual inspection of the average estimated interaction effect sizes actually decreased in size from first (Δd = -.46) to last (Δd = -.19), although confidence intervals of these estimates did overlap.

To interpret the interactions it is important to recall (see Fig. 3~~2~~) that the *overall* ostracism effects are relatively large and operated similarly at both levels of the cross-cutting moderator variable. Moreover, when we compared the mean effects of the moderator variable *within* the two possible levels of ostracism factor (i.e., ostracized or include), results indicate a relatively weak *positive* effect within the ostracism level and a relatively weak *negative* effect within the inclusion level. To further explain the implication of the findings it may be fruitful to consider an example in which participants are ostracized or included by either an outgroup or an ingroup. In such a setting, our findings would thus suggest that the relative effect of ostracism compared to inclusion (i.e., the ostracism effect), is similar for both outgroup *and* ingroup conditions. Moreover, if one compares the effect of group status (outgroup vs. ingroup), one would predict that those ostracized by outgroup members would slightly benefit whereas those included by ingroup members would slightly be harmed. Taken together, these contrasts support the robustness of the ostracism effect.[7~~6~~]

Structural Aspects of Cyberball and Different Dependent Variables

The secondary analyses confirmed that the overall findings generalize to a large extent across structural aspects, sampling aspects and type of dependent variable.

Does gender of participants matter?

Previous research provided evidence for a difference in the ostracism effect across genders [18]. Our results indicated that, contrary to this, proportions of males and females did not significantly predict the mean effect size. In our coded studies, the mean proportion of males was approximately 39% (observed range: 0-100%).

Does age of participants matter?

Whereas previous research has indicated increased sensitivity to ostracism in younger age groups [19], we failed to find moderation of ostracism effects by mean age of the study samples. Coded studies had a mean sample age ranging from 10 through 32.5 years, with an average of approximately 20.5 years. This indicates that most of the research with Cyberball has been done on young adults, with relatively few or no studies investigating children, middle-aged participants, or senior citizens. More research could focus on specific (individual-level) age moderation of ostracism.

Does culture or country matter?

We found no indication that culture predicted the average effect size. In our coded studies, approximately 52% were from the United States, 45% from other Western countries (e.g., Australia, the Netherlands, Germany),~~)~~ and 3% from Asian countries. Our analyses used the United States as reference category. We note that the low prevalence of Asian countries might cause a lack of power~~,~~ and that we cannot definitively state there is no difference between Western and Asian responses to ostracism. We can state that there is no systematic difference in the ostracism response for Western countries and the United States.

Does number of players matter?

In the studies included in this meta-analysis, approximately 89% of the studies used the three-player version of Cyberball and 11% used the four-player version of Cyberball. Average ostracism effects differed between these subsets, with smaller predicted effects in the four-player setting, but we are hesitant to interpret this due to a nonsignificant omnibus test for the predictive model (see ‘Composition’ in the results section). Preferably, this moderator of the ostracism effect in Cyberball should be subject to further work in which the number of players is experimentally varied.

Does number of throws or length of the study matter?

We considered the length of Cyberball in two ways. We coded the number of ball tosses and estimated the length of the study. Of the coded studies, 60% used 30 throws, 11% used 40 throws, 8% used 20 throws, 4% used 60 throws, and 2% for both 15 and 24 throws. Other categories ranging from 10 through 200 make up the remaining percentages, each making up 1%. Only 2 out of 120 studies were estimated to last longer than 5 minutes. Our results indicated the mean ostracism effect was *not* reliably predicted to be different across different lengths of the study or the different number of total throws in the omnibus test. The single meta-regression on ball tosses suggested it may predict the effect size of the first measure. As above, we are hesitant to interpret this, but do note that increasing ball tosses may be more associated with a diffused ostracism effect than with an increased ostracism effect.

Does type of dependent variable matter?

Secondary analyses also showed that the majority of the results were robust across subsets of dependent measures and the overall set of dependent measures (see Fig. 3~~2~~). Exceptions were interpersonal measures showing relatively weaker ostracism effects on the first measure when compared to the other subsets. This suggests that psychological effects of ostracism are large, but that this effect might be smaller for interpersonal behaviors. On top of this, interpersonal measures also show more moderation, suggesting that interpersonal behaviors caused by ostracism are more easily moderated by cross-cutting factors. Additionally, we estimated interactions for the measure subsets interpersonal (i.e., measures relating to others), intrapersonal (measures relating to the self), fundamental needs, model (i.e., first measure is reflexive and last measure is reflective),~~)~~ and an overlap of the latter two subsets. For all but two, these subsets showed that measures taken at the first time point were moderated more strongly than the measures taken last. Finally, the analyses including only fundamental needs showed that moderation was larger at the last time point, when compared to the first time point. This result is crucial, as Williams [11] specifically predicted this pattern for fundamental needs.

Williams's Model of Ostracism: Supported or Not?

Regarding the test of Williams's [11] model, there are several important observations and limitations. First, Williams proposed fundamental need threat as a result of even a brief episode of ostracism. This was supported by the meta-analysis. Moreover, moderation is predicted to occur in the reflective stage, when the context and meaning of the ostracism event can be appraised. This was also supported in the present meta-analysis. The final stage of

Williams’s model—resignation—is outside the aims of the present meta-analysis, because it requires long-term exposure to ostracism.

The proposition that appears to lack support from this meta-analysis is that reflexive reactions to ostracism are more resistant to moderation than reflective reactions. Across the board, our results indicate there is more moderation of ostracism effects on the first time point than on the last time point. However, there are two limitations to this conclusion. Firstly, Williams specifically refers to physiological, online, or immediate retrospective reports to assess reflexive reactions. In many instances in this meta-analysis, the first reaction is not isomorphic with reflexive measures. Anything taken after the game, or assessed by wording indicating present state (rather than the participants’ state during the game), is not assumed to be reflexive, nor predicted to be resistant to moderation. Secondly, Williams’s proposition is restricted to fundamental needs only. Indeed, our specific analyses involving only studies that employed measures of immediate and delayed fundamental need satisfaction corroborated the model prediction that there is more moderation on the last time point, than on the first time point.

Because of this quantitative difference in moderation across measures, we encourage direct testing of this time difference in moderation as predicted by Williams [11], just as the study by Bernstein and Claypool [39] was a direct, experimental test of a finding by Gerber and Wheeler [14]. However, the mean size of the interaction effect in out meta-analysis was quite small, raising power issues for future studies. Using our estimated interaction effects to determine sample size under a power of .8, a sample size of 2186 would be necessary to have sufficient power on both time points.[8~~7~~] Note that the mean sample size in full factorial designs in our meta-analysis is 110, showing that the mean power in these studies is .08 to detect an *interaction* at the last time point (notably, power for the standard ostracism effect is highly sufficient in the included studies, due to the large effect). A large Mechanical Turk study is feasible and could provide the sample needed. Additional ways of increasing power are by reducing error on the measurements by using validated psychometric scales.

Changes to the need-threat model of ostracism

As a result of our findings, we suggest that the temporal need-threat model of ostracism should be modified. Firstly, it should be recognized that there is potential for moderation in the reflexive stage, where immediate measures of impact tap into participants' reactions during the game. If factors can reduce physical pain and distress, like for instance acetaminophen [40][9]~~[8]~~ or transcranial magnetic stimulation [41], or if certain populations are less likely to feel pain (e.g., those higher in schizotypal personality disorder [42,43]), then we would also expect moderation of immediate measures of distress. Secondly, our results may suggest important issues related to the timing of measuring ostracism effects by way of the ordinal differences. Specifically, time passed after the ostracism episode occurred is likely to affect the extent immediate distress measures will be subject to moderation. For example, if researchers wait long enough before administering the immediate need satisfaction measures (e.g.,~~.~~ "playing the game made me feel insecure"), it becomes more likely that all participants will have recovered from the negative impact of ostracism, thus resulting in a homogeneous (and highly satisfied) between-group result. Thus, differences in recovery from ostracism based upon social-situational factors and/or personality differences, if any, occur somewhere between initial pain and final recovery. It is difficult to predict exactly when that time period is. Zadro et al. [44] report delayed recovery by those high in social anxiety 45-minutes later. Other studies show full recovery within 5-10 minutes. Future research needs to examine the time course more carefully, to determine if and when moderation occurs in delayed measures.

Limitations

Within the current meta-analysis there are several limitations. One potential limitation is that our testing of differences between first and last measure was indirect. We compared confidence intervals to evaluate whether the effects were different. A direct test would provide more conclusive evidence on whether or not the effects are indeed equal or different across the first and last measurements. Note, however, that a direct test requires correlations between the measurements for every study, every condition, and every type of different variable. This information was not given in the vast majority of the papers and we anticipated that a direct request for such information would suffer from the problem of low response rates [45] which would in turn lower the sample size of the meta-analysis and thus the ability to effectively test our hypotheses.

A second potential limitation is that the random (non-systematic) heterogeneity in the effect sizes poses a problem for the power of finding moderator effects [25]. This could pose the problem that several of the non-effects found are actually there, but not detected (Type II errors). However, our subset analysis of typical Cyberball studies – ~~i.e.,~~ 3 players games involving 30 ball tosses, lasting less than five minutes, with immediate fundamental need satisfaction as dependent variable - still showed substantial variability in the effect sizes: I^2 = 83%. This indicates that the effects are quite variable to begin with~~,~~ and makes it unlikely that the overall effects are misrepresented.

Also, we did not observe that our estimation of time predicted the ostracism effect on the last measure. This null-effect may be a reality but could also be caused by the fact that the (random) heterogeneity in the effect sizes may have been too large to find moderation by time. This cannot be counteracted in the current dataset and remains a limitation. Second, imprecise reporting of the measures in the papers may have led to inaccurate time estimations. To counteract this imprecise reporting of measures, authors could be contacted, but this also poses new problems (i.e., nonresponse, or authors might not be willing to admit that measures were left out in the paper [46]).

Importantly, we did observe that the confidence intervals of both the first and last measure did not overlap, suggesting that there is a ~~qualitative~~ difference in effect size between first and last measure. The question then is whether this difference is indeed caused by time of measurement or in part caused by the type of measurement used across the two different time points. This explanation can be addressed by inspecting whether~~creating a difference index in which~~ the composition of~~difference in dependent~~ measures is different across time points. On~~at~~ the first ~~and second time point are inspected by creating a difference index (i.e., coded value on first~~ measure 0.84 was intrapersonal self-report, 0.02 was intrapersonal physiological, 0.01 was intrapersonal other, 0.08 was interpersonal anti-social, 0.03 was interpersonal pro-social, and 0.01 interpersonal other. On the ~~minus coded value on~~ last measure 0.79 was intrapersonal self-report, 0.04 was intrapersonal physiological, 0.02 was intrapersonal other, 0.05 was interpersonal anti-social, 0.08 was interpersonal pro-social,~~)~~ and 0.01 was interpersonal other. This shows that ~~regressing the index on~~ the different ~~observed effect sizes in a meta-regression. Doing this for the standard ostracism effect on the last measure, showed no significant predictive effect of this difference (*b* = -0.03, *p* = .531), indicating that the difference in estimated effects is not driven by difference in measures on the first and last time point. Also, inspecting whether the~~ types of dependent variables~~measures used across all studies are different, and not the difference within a study, shows that these~~ are similarly distributed across time points (maximum discrepancy of 4.9 percentage points). Substantive differences in proportions of measures across time points are minimal and thus form an unlikely driving force for our findings.

A third limitation is that this paper only summarized the results of the measures included in the studies. However obvious this might be, it should be pointed out, because the validity of the conclusions are reliant on the validity of the measures. Most prominently represented in the current meta-analysis are the fundamental need measures, which have no proper psychometric validation up-to-date, notwithstanding their wide use. Other kinds of included measures possibly also lack proper validation~~,~~ and one has been openly criticized (e.g., the Hot Sauce aggression paradigm [47]). ~~We note that results in this paper are conditional on that these measures are valid.~~

Conclusion

Our meta-analysis of 120 Cyberball studies extends the temporal need-threat model of ostracism. We observed that the average effect size approaches 1.5 standard deviations and that this average effect size is not affected by the composition of the sample used (i.e., age, gender, country of origin) nor by structural aspects of the game (i.e., number of ball tosses, duration, players). We also observed that findings are relatively robust across the typical dependent variables that are used in Cyberball and that the overall effect size decreases from first to last measure. Importantly, we also observed that first measures can be moderated by cross-cutting variables and that only fundamental needs measures show stronger moderation for the last measures as opposed to the first measure taken in the studies. The moderation analyses by cross-cutting variables also revealed that the interaction effects sizes are considerably smaller than the direct inclusion vs. ostracism effect size. This revealed that the typical Cyberball study has enough power to detect main effects, but should substantially increase sample size to study theoretically relevant interactions. Intriguingly, we also observed that effect sizes were rather heterogeneous even when we limited our analysis to a very homogenous subset of studies. This indicates that there are potentially relevant moderators that have yet not been discovered. We invite fellow researchers to reanalyze our data (osf.io/ht25n) and test new hypotheses, and to further expand our knowledge of ostracism with Cyberball.

Footnotes

1. The direct link: https://osf.io/ht25n/

2. It has been updated since, but the list that was used can be found on the Open Science Framework, see Footnote 1.

3. Oaten, Williams, Jones and Zadro [48] was applicable, but was excluded due to being an outlier with respect to effect size (*ds* > 15). See also Gerber and Wheeler (2009; p. 473): "*One study (Oaten, Williams, Jones, & Zadro, 2007) had need effect sizes that were clear outliers (effect sizes were 5–7 standard deviations above the means)* […and…] *were excluded from the analyses.*"

4. Due to the dependency between the standardized effect size and the standard error, we also ran an alternative version of the Egger's test that regresses on 1/N. These analyses yielded highly similar results.

~~4.~~5. Because fundamental needs showed effects in the theorized direction, we explored this further by overlapping the subset of fundamental need measures with the model definition of immediate and delayed (i.e., whether the measures related to feelings during or after the Cyberball game). Estimated interactions for this selection were Δd = -0.37, 95% CI [-0.60, -0,14] (k = 29) and Δd = -0.13, 95% CI [-0.53, 0.27] (k = 8) for the first and last measure, respectively. So in this particular subset of studies that use immediate or delayed fundamental needs measures, results are not in line with Williams's [11] prediction. The reported fundamental need selection can be specified even further to only include studies that explicitly focus on composite need satisfaction as typically defined by Kip Williams. Such a selection again provides support for the hypothesis that immediate fundamental need satisfaction is less moderated, Δd = -0.18, 95% CI [-0.47, -0.11] (k = 15), than delayed need satisfaction,

Δd = -0.93, 95% CI [-1.67, -0.19] (k = 3). Note, however, that such a selection is based on 3 studies for delayed measures.

~~5.~~6.We also conducted individual meta-regressions for each of the structural- and sampling variables. These individual analyses yield similar results as the overall analyses. We again observed that four players are less hurt by ostracism than three players (b = .84, SE = .28, p = .003) on the last measure. What is new is that we also observed that number of ball tosses affected the effect size (b = .02, SE = .01, p = .046) on the first measure. This showed that increasing the number of ball tosses decreases the negative impact of ostracism. Taken together this suggests that the impact of ostracism is diffused when it is the result of more players and more ball tosses compared to fewer players and fewer balls tosses.

~~6.~~7.It is important that the simple effects in Fig. 3~~2~~ are averaged over studies, thus potentially subject to Simpson's paradox.

~~7.~~8. We used G*Power 3.1.7 to calculate this between-subjects interaction effect (F-test, fixed effects, .8 power); with k = 4 and the smaller interaction (last time point; numerator $df = k – 1$). The effect size Δd was transformed in to f by means of $\sqrt{[d^2/(2k)]}$, resulting in f = .0707.

~~8.~~9.DeWall et al. was not included in the meta-analysis, because we were not able to retrieve all information.

Appendix

All formulae reported below originate from the chapter by Michael Borenstein (2009). Hedges' *g* was calculated as

$$g = d\left(1 - \frac{3}{4df_w - 9}\right)$$

where *d* is the standardized main effect and df_w is the number of conditions minus 1.~~,~~ For the standardized interaction effect *d* was calculated as

$$\Delta d = \frac{(\bar{X}_{11} - \bar{X}_{12}) - (\bar{X}_{21} - \bar{X}_{22})}{s_p}$$

where the first term in the numerator~~nominator~~ is the ostracism effect and the second term is the ostracism effect in the moderator conditions. When transformed to a squared correlation coefficient, this~~This~~ Δ*d* corresponds to the partial eta-squared of the interaction. Sampling variance of *g* was calculated by multiplying the sampling variance of *d* by the squared correction factor, that is

$$s_g^2 = \left(1 - \frac{3}{4df_w - 9}\right)^2 s_d^2 \; s_g = \left(1 - \frac{3}{4df_w - 9}\right)^2 s_d$$

where the sampling variance of the interaction was calculated as the sum of the sampling variances of both the simple main effects.

Formatted: Font: (Default) Times New Roman, (Asian) +Body Asian (SimSun), 12 pt, Lowered by 16 pt

Acknowledgements

We would like to kindly thank all researchers who sent in their unpublished studies and helped us collect all the data we needed. We would also like to thank Andrew Hales for his constructive comments and feedback. The preparation of this article was supported by grant number 016-125-385 from the Netherlands Organization for Scientific Research (NWO) awarded to Jelte M. Wicherts and by the NSF under Grant #BCS-1339160 awarded to Kipling D. Williams.

References

References marked with an asterisk indicate studies included in the meta-analysis.

1. *Williams KD, Cheung CK, Choi W (2000) Cyberostracism: effects of being ignored over the Internet. J Pers Soc Psychol 79: 748–762.
2. Baumeister RF, Twenge JM, Nuss CK (2002) Effects of social exclusion on cognitive processes: Anticipated aloneness reduces intelligent thought. J Pers Soc Psychol 83: 817–827.
3. Nezlek JB, Kowalski RM, Leary MR, Blevins T, Holgate S (1997) Personality moderators of reactions to interpersonal rejection: Depression and trait self-esteem. Personal Soc Psychol Bull 23: 1235–1244.
4. Craighead WE, Kimball WH, Rehak PJ (1979) Mood changes, physiological responses, and self-statements during social rejection imagery. J Consult Clin Psychol 47: 385–396.
5. Leary MR, Kowalski RM, Smith L, Phillips S (2003) Teasing, rejection, and violence: Case studies of the school shootings. Aggress Behav 29: 202–214.

6. *Lustenberger DE, Jagacinski CM (2010) Exploring the Effects of Ostracism on Performance and Intrinsic Motivation. Hum Perform 23: 283–304.

7. *Carter-Sowell AR, Chen Z, Williams KD (2008) Ostracism increases social susceptibility. Soc Influ 3: 143–153.

8. *Van Beest I, Carter-Sowell AR, van Dijk E, Williams KD (2012) Groups being ostracized by groups: Is the pain shared, is recovery quicker, and are groups more likely to be aggressive? Gr Dyn Theory, Res Pract 16: 241–254.

9. Baumeister RF, Leary MR (1995) The need to belong: desire for interpersonal attachments as a fundamental human motivation. Psychol Bull 117: 497–529.

10. *Ijzerman H, Gallucci M, Pouw WTJL, Weiβgerber SC, Van Doesum NJ, et al. (2012) Cold-blooded loneliness: social exclusion leads to lower skin temperatures. Acta Psychol (Amst) 140: 283–288.

11. Williams KD (2009) Ostracism: a temporal need-threat model. Adv Exp Soc Psychol 41: 275–314.

12. Haselton MG, Buss DM (2000) Error management theory: a new perspective on biases in cross-sex mind reading. J Pers Soc Psychol 78: 81–91.

13. Blackhart GC, Nelson BC, Knowles ML, Baumeister RF (2009) Rejection elicits emotional reactions but neither causes immediate distress nor lowers self-esteem: a meta-analytic review of 192 studies on social exclusion. Pers Soc Psychol Rev 13: 269–309.

14. Gerber J, Wheeler L (2009) On Being Rejected: A Meta-Analysis of Experimental Research on Rejection. Perspect Psychol Sci 4: 468–488.

15. Cacioppo S, Frum C, Asp E, Weiss RM, Lewis JW, et al. (2013) A Quantitative Meta-Analysis of Functional Imaging Studies of Social Rejection. Sci Rep 3.

16. Rotge J-Y, Lemogne C, Hinfray S, Huguet P, Grynszpan O, et al. (2014) A meta-analysis of the anterior cingulate contribution to social pain. Soc Cogn Affect Neurosci: nsu110.

17. *De Waal-Andrews W, van Beest I (2012) When you don’t quite get what you want: psychological and interpersonal consequences of claiming inclusion. Pers Soc Psychol Bull 38: 1367–1377.

18. *Hawes DJ, Zadro L, Fink E, Richardson R, O’Moore K, et al. (2012) The effects of peer ostracism on children’s cognitive processes. Eur J Dev Psychol 9: 599–613.

19. *Pharo H, Gross J, Richardson R, Hayne H (2011) Age-related changes in the effect of ostracism. Soc Influ 6: 22–38.

20. Hofstede G (1980) Culture’s consequences: International differences in work-related values. London, UK: Sage.

21. *Van Beest I, Williams KD (2006) When inclusion costs and ostracism pays, ostracism still hurts. J Pers Soc Psychol 91: 918–928.

22. *Zadro L, Williams KD, Richardson R (2004) How low can you go? Ostracism by a computer is sufficient to lower self-reported levels of belonging, control, self-esteem, and meaningful existence. J Exp Soc Psychol 40: 560–567.

23. Hunter J, Schmidt F (1990) Dichotomization of continuous variables: The implications for meta-analysis. J Appl Psychol 75: 334–349.

24. MacCallum RC, Zhang S, Preacher KJ, Rucker DD (2002) On the practice of dichotomization of quantitative variables. Psychol Methods 7: 19–40.

25. Hedges L V, Pigott TD (2004) The power of statistical tests for moderators in meta-analysis. Psychol Methods 9: 426–445.

26. Williams KD, Jarvis B (2006) Cyberball: A program for use in research on interpersonal ostracism and acceptance. Behav Res Methods 38: 174–180.

27. Smits IAM, Dolan C V, Vorst H, Wicherts JM, Timmerman ME (2011) Cohort differences in Big Five personality factors over a period of 25 years. J Pers Soc Psychol 100: 1124–1138.

28. *Gonsalkorale K, Williams KD (2007) The KKK won’t let me play: ostracism even by a despised outgroup hurts. Eur J Soc Psychol 37: 1176–1186.

29. Viechtbauer W (2010) Conducting meta-analyses in R with the metafor package. J Stat Softw 36: 1–48.

30. R Core Team (2013) R: A language and environment for statistical computing. Available: http://www.r-project.org/.

31. Hedges LV (1981) Distribution theory for Glass’s estimator of effect size and related estimators. 6: 107–128.

32. Viechtbauer W (2005) Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. J Educ Behav Stat 30: 261–293.

33. Light RJ, Pillemer DB (1984) Summing up: the science of reviewing research. Cambridge, MA: Harvard University Press.

34. Bakker M, Van Dijk A, Wicherts JM (2012) The rules of the game called psychological science. Perspect Psychol Sci 7: 543–554.

35. Egger M, Smith GD, Schneider M, Minder C (1997) Bias in meta-analysis detected by a simple, graphical test. BMJ 315: 629–634.

36. Sterne JAC, Egger M (2005) Regression Methods to Detect Publication and Other Bias in Meta-Analysis. In: Rothstein HR, Sutton AJ, Borenstein M, editors. Publication bias in meta-analysis. Chichester: John Wiley & Sons.

37. Cohen J (1988) Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum.

38. Schenker N, Gentleman JF (2001) On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals. Am Stat 55: 182–186.

39. *Bernstein MJ, Claypool HM (2012) Not all social exclusions are created equal: Emotional distress following social exclusion is moderated by exclusion paradigm. Soc Influ 7: 113–130.

40. DeWall CN, MacDonald G, Webster GD, Masten CL, Baumeister RF, et al. (2010) Acetaminophen reduces social pain: behavioral and neural evidence. Psychol Sci 21: 931–937.

41. *Riva P, Romero Lauro LJ, Dewall CN, Bushman BJ (2012) Buffer the pain away: stimulating the right ventrolateral prefrontal cortex reduces pain following social exclusion. Psychol Sci 23: 1473–1475.

42. *Wirth JH, Lynam DR, Williams KD (2010) When social pain is not automatic: Personality disorder traits buffer ostracism’s immediate negative impact. J Res Pers 44: 397–401.

43. Lautenbacher S, Krieg J-C (1994) Pain perception in psychiatric disorders: A review of the literature. J Psychiatr Res 28: 109–122.

44. *Zadro L, Boland C, Richardson R (2006) How long does it last? The persistence of the effects of ostracism in the socially anxious. J Exp Soc Psychol 42: 692–697.

45. Wicherts JM, Borsboom D, Kats J, Molenaar D (2006) The poor availability of psychological research data for reanalysis. Am Psychol 61: 726–728.

46. LeBel EP, Borsboom D, Giner-Sorolla R, Hasselman F, Peters KR, et al. (2013) PsychDisclosure.org: Grassroots Support for Reforming Reporting Standards in Psychology. Perspect Psychol Sci 8: 424–432.

47. Ritter D, Eslea M (2005) Hot Sauce, toy guns, and graffiti: A critical account of current laboratory aggression paradigms. Aggress Behav 31: 407–419.

48. Oaten M, Williams KD, Jones A, Zadro L (2008) The effects of ostracism on self-regulation in the socially anxious. J Soc Clin Psychol 27: 471–504.

49. Borenstein M (2009) Effect sizes for continuous data. In: Cooper H, Hedges L V., Valentine JC, editors. The handbook of research synthesis and meta-analysis. New York, NY: Russell Sage Foundation.

Supporting information

S1 File. Data package. Contains data and the R analysis script.

S2 File. Full reference list meta-analysis studies. Contains the full reference list of the studies included in the meta-analysis.

S3 File. Scatterplot of the effects in hypotheses 1 and 2 and estimated time.

S4 File. Figure 3~~2~~ subset lists. Contains the lists of what studies that were in the meta-analysis are included in computing the effects for the different panels.

~~**S4 File. PRISMA flow diagram.**~~

Page 1

Ilja van Beest
Faculty of Social and Behavioral Sciences
Tilburg University
P.O. Box 90153
5000 LE Tilburg
The Netherlands

Prof. Dr. Nico van Yperen
Academic Editor, PLOS ONE
Rijksuniversiteit Groningen

Manuscript ID PONE-D-15-02806

Dear Prof Dr. van Yperen,

Thank you for your kind words about our work and for the opportunity to submit a revised version of our paper “The Ordinal Effects of Ostracism: A Meta-Analysis of 120 Cyberball Studies" to PLOS ONE. As you recommended, we have carefully considered each of the comments made by the reviewers, paying special attention to those highlighted by you in your letter. A detailed overview of our revisions is included below. For your convenience we copied the three reviews and added a detailed description of how we made the appropriate changes immediately below each comment.

We believe that the changes we made have substantially improved the manuscript and made our contribution stronger. We warmly thank you for your help in achieving this and look forward to your final decision.

Kind regards, also on behalf of Chris Hartgerink, Jelte Wicherts, and Kip Williams

Ilja van Beest

Page 2

Reviewer #1: The authors have conducted a meta-analysis of studies using the Cyberball game, which manipulates the degree of social inclusion versus ostracism experienced by participants. Particular focus in this meta-analysis is on the immediate and delayed effects of the experimental manipulation and on examining whether immediate or delayed effects are more susceptible to the moderating influence of other factors.

In general, I think that PLoS One is an appropriate outlet for this meta-analysis and I would support its publication. However, below I list a number of general issues, concerns, comments, and appeals for clarification that I think the authors need to address first.

General Issues / Concerns / Comments:

#1

page 7: Author predictions were used to determine how an interaction should be coded. Was there always a clear prediction given by authors so that this decision could be made unambiguously? If not, was the intercoder reliability of these assessments measured? Aside from this, we know that some 'predictions' are actually generated post-hoc, after the results have become available. That is a limitation that should be acknowledged.

Answer

Of the 120 studies that were investigated, 52 studies contained an interaction. The prediction in these 52 studies, was based on the explicit prediction of the authors of the manuscript. Moreover, the first authors (Chris and Ilja) checked and discussed each paper until consensus was reached. We did not record these discussions and intercoder reliability cannot be assessed. We did provide a case by case description of all studys on OSF.

We acknowledge that the predictions of the primary studies could be post-hoc and this is now acknowledged in the revised manuscript. We now say

A potential limitation of our decision to follow the prediction of the authors is that the predictions may have been generated post-hoc on the basis of observed outcomes.

on page 7 line 135.

#2

page 10, line 208: In studies with more than one additional factor (besides the ostracism factor), the authors "collapsed effect sizes across the factor that authors expressed least interest in." I can imagine that this decision cannot always be made with 100% certainty. Did the authors attempt to estimate the intercoder reliability of these assessments?

Answer

Seventeen of the 52 studies with a cross-cutting variable involved designs that were more complex than the 2x2 design. In these studies, the selection decisions were jointly made by Chris Hartgerink and Ilja van Beest. Intercoder reliability was not assessed.

#3

page 10, line 224: I know from personal experience that one of the last things that authors of a meta-analysis want to hear is: Your search is outdated. Indeed, a meta-analysis may go through through several (re)submission rounds before being accepted/published and the date of the search then increasingly falls further behind. There is in principle no need to demand an update, so I will not insist. However, are the authors aware of any additional studies that have become available after their search was concluded?

Answer

We agree that the search for additional studies is time-consuming and that one should always chose a moment to stop updating the database. Nevertheless, we conducted another search in Web of Knowledge for Cyberball studies, which resulted in 71 hits for 2013-2015 (searched on March 17, 2015). After inspecting which of these studies would have met our inclusion criteria, 29 remained after our previous end date (April 2013). These 29 references are available here (EndNote format). Of these 29, we already included 2 studies that were not published when we collected them, and 14 contained a cross-cutting variable. Given the current size of the database and the sample sizes in these new studies, we do not expect them to significantly change any of our core conclusions. Hence, we decided not to redo all of the analyses using this updated database.

#4

page 13: I am wondering about the selection/coding of the first and last measure. Was there never any ambiguity regarding the order in which instruments were administered? Also, if authors said that they used measures X, Z, and Z after the game, the actual order may have been different.

Answer

We based the coding of the first- and last measure on the information presented in the paper describing the primary study. This information was straightforward and we did not encounter ambiguity regarding the order in which the instruments were administered. We acknowledge that people may have included more measures than reported and that unreported measures remain unaccounted for, such that the estimate for time between the first and last is a crude one. In other words, we could not get better information than that reported in the paper, which is why we retain the information reported in the paper as the most viable situation.

#5

page 13: First and last measures were classified into four categories (interpersonal, intrapersonal, fundamental needs, or model correspondence). So, if I understand the authors correctly, first a measure was chosen as being first/last and then this classification was made (so there is always exactly one first measure and if the study applied multiple/delayed assessments, there is always exactly one last measure). Can the authors please confirm/clarify this?

Answer

We hereby confirm that every study contained a first measure and if present, a last measure. Table 2 illustrates this, where some studies do not contain an effect on the last measure.

#6

page 13: Also, does that imply that the first measure may have assessed, for example, intrapersonal effects, while the last measure may have assessed, for example, interpersonal effects? Or in other words, is it possible that the effect size estimates in Table 2 (d_T1 and d_T2, and similarly, Delta-d_T1 and Delta-d_T2) actually reflect different measurement types? This needs to be clarified, since this has major implications for the interpretation of the results reported on pages 25 to 27.

Answer

Yes, this is correct. Figure 2 separates the effects per type of measure and shows that results are consistent across the different types of dependent variables, except for interpersonal behavior (as mentioned in the text).

#7

page 18, line 350: I am not sure if "standardized simple effects across the ostracism factor" is appropriate terminology here (and elsewhere in the paper). In a two-way factorial design, a "simple effect" is the effect of one factor *within* one of the levels of the other factor. So, if that other factor has two levels, then there would be two simple effects. That would apply to each time point, so in a 2x2 design with multiple measures (one of which is the first and one is the last measure), there would be 4 (not 2) simple ostracism effects. However, if I understand the authors correctly, they are not computing simple effects here, but marginal/main effects for the first and for the last measure (i.e., the difference between the ostracism and inclusion levels averaged over any other factors). Please clarify this (and the terminology throughout the manuscript).

Answer

We did intend simple effects, as we calculated four simple effects for the ostracism factor (one in the moderated conditions, one in the non-moderated conditions, for both first and last measure). The reviewer refers to the set of 52 studies where a second factor is included, where we calculated the simple effect of ostracism within the non-moderated level. We clarified this in the revised manuscript. Specifically, we now write:

Standardized effects were calculated across the ostracism factor, where the 52 studies with a cross-cutting variable were included as a simple effect of ostracism within the non-moderated level.

On page 18 ~ line 349. Additionally, we deleted the following to prevent confusion (lines 355-356):

Non-factorial studies delivered only simple effects for the first and last measure, and no interactions

#8

page 18: The description of the interaction effect given here (and on the previous pages and also the appendix) suggests that moderators of the ostracism effect can take on only two values/levels. However, was that always the case?

Answer

Moderator factors could include more levels, in which case we selected the two conditions that were the farthest apart in design. For example, if a study included an ostracism factor (included or ostracized) and a players factor (3, 5, 10, 15 players) as a moderator, we used the 3 and 15 player levels. Selection based on the factorial levels occurred in 10 studies. We mention this number in the text of the revised manuscript (page 18 line 359)

Table 2:

#9

1) I see many rows where "First author" and "Year" is identical. Can the authors explain how this arises?

Answer

We thank the reviewer for this comment. The reason is that papers may contain multiple studies. To clarify this, we now added a note.

Multiple rows for the same first author and year is possible due to multiple studies across papers.

#10

2) In the table notes, the authors write: "Non-integer Ns arise from division of full sample N for included conditions, appropriate due to random assignment." I don't understand what the authors mean by this (and I could find no further discussion of this in the paper).

Answer

Ns of for example 12.333 arise from a 3-condition design, where random assignment was used. If N per condition was not given, we divide total N (e.g., 37) by the number of conditions (3) to come to a condition N estimate. To clarify we added an example in the table note:

(e.g., two conditions out of 3, when sample is 56: (56 / 3) × 2 = 37.333)

#11

3) It appears that multiple estimates are often obtained from the same study. Given that "N" differs for these rows, these effects seem to be based on different samples, so within a particular study, the estimates may be independent. However, that still does not preclude the possibility that multiple estimates obtained from the same study are more similar to each other than estimates obtained from different studies. In other words, the data seem to have a multilevel structure, which would imply the need to employ an appropriate multilevel meta-analysis model that accounts for such dependencies (e.g., by adding a random effect at the study level to the current model).

Answers

The reviewer notes that the data may be interdependent within an analysis; this is incorrect. Effects that go into the same meta-analysis are independent (i.e., one effect per study): every row is an independent study, which also explains the difference in N. However, the reviewer is *correct in stating that from one paper multiple independent studies can be included. This multilevel modeling is therefore not necessary.*

#12

page 25: I assume the authors applied the version of Egger's regression test that relates the effect size estimates to their standard errors. For standardized mean differences, the standard error depends on the size of the effect, which can cause spurious associations especially when effects are large. Similar deficiencies of the test have been observed when using effect size measures based on dichotomous data (e.g., risk/odds ratios or risk differences). For a more appropriate version of the test, the authors should use some measure of precision that does not depend on the size of the effect, the obvious choices being the sample size, the inverse sample size, or square-root transformations thereof.

Answer

As requested by the reviewer, we conducted these regression tests with 1/N as predictor. Results are the same as the Egger's test with standard error as predictor and is therefore not adjusted further in the manuscript. We include a footnote in the methods section of the manuscript that reads:

Due to the dependency between the standardized effect size and the standard error, we also ran an alternative version of the Egger's test that regresses on 1/N. These analyses yielded highly similar results.

#13

page 25: Coding the estimated time between exclusion and the moment at which the last measure was taken in *seconds* seems artificially precise. Did the authors calculate the intercoder reliability for these estimates based on independent coders? Also, please rescale this moderator into some larger units (e.g., minutes) which avoids the extremely small coefficient (.0001). In addition, since this is one of the primary hypotheses tested in the paper, please provide a scatterplot of the time variable against the effect size estimates.

Answer

Following the suggestion of the reviewer we rescaled the time estimate into minutes. The results have been adjusted accordingly.

Also note that the time estimation was based on the number of items times the six second rule, plus any additional time mentioned in the paper. This information was readily available in all manuscripts although we acknowledge that it is possible that not all dependent variables were disclosed in a paper describing the study (see also our answer reviewer 1, #4). As mentioned, in the 68 studies without cross-cutting variable were coded by Chris Hartgerink, the 52 with a cross-cutting variable were coded by both Chris Hartgerink and Ilja van Beest. Consensus was readily reached and we did not collect quantitative information to calculate intercoder reliability.

Following the suggestion of the reviewer, we now provide scatterplots of time versus effect (simple and interaction on timepoint two) in the Supplemental Materials of the revised manuscript.

#14

page 27: Same issues apply here. I cannot imagine that two independent coders would ever come to the exact same assessment when coding time in seconds. Also, please rescale time to avoid the overly small coefficient. And please provide a scatterplot.

Answer

See answer (reviewer #1, answer #13).

#15

page 28 and Figure 2: As far as I can tell, here the authors are indeed talking about simple effects (e.g., "the between-subjects effect of being ostracized with no moderator present, whereas moderated ostracism effect refers to being ostracized with a moderator present"). Earlier, the authors also talked about "simple effects" (which I think are actually main effects -- see my earlier comment -- but maybe I am misunderstanding what the authors did). Please clarify this.

Answer

See answer (reviewer #1 answer #7).

#16

Also, if I understand Figure 2 correctly, I would assume then that the *difference* between, let's say, the points for "All" in panels (1) and (2) is equal to the *difference* between the points for "All" in panels (5) and (6) (since the difference between the two simple effects for factor A within the two levels of factor B must be equal to the difference between the two simple effects for factor B within the two levels of factor A). However, visual inspection suggests that this may not the case. Can the authors clarify?

Answer

We are not sure whether we understand the question. It seems that the reviewer postulates that the difference in the simple effects for ostracism on the different moderator levels is supposed to be equal to the difference in simple effects for the moderator levels on the ostracism levels. Below we provide an example that this would be incorrect and that simple effects do differ.

	N-mod	mod
Ostr	5	7
Incl	2	3

In this case, the simple effect of ostracism is 5-2 = 3 for the non-moderator level and 7-3 = 4 for the moderated level. For the simple effect of moderator within the ostracism level, 5-7 = -2 and within the included level 2-3 = -1. Correspondingly, simple effects all differ and are not required to be equal, as the reviewer proposes.

#17

page 30, line 514: "Model indicates" -- which model?

Answer

The model pertained to a subset included throughout the analyses. To avoid confusion we rewrote the note under table 3 to read similar to Figure 2

The subset labeled “All” contains all measures. The subset labeled “Fundamental” contains only fundamental need measures. The subset labeled “Intrapersonal” contains all intrapersonal measures. The subset labeled “Interpersonal” contains all interpersonal measures. The subset labeled “Model” contains those where first measures is immediate and last measure is delayed. See Supplement S4.

On page 28 this was clarified under the heading Measures, where the subsets are named.

#18

page 30, lines 515-516: I don't understand what the authors mean by "listwise deletion for equal ks across time points". Please clarify.

Answer

To clarify what we mean by listwise deletion we adjusted the sentence as follows:

Listwise deletion ensures that estimates are made on full rows in the data. Listwise deletion was applied in all the subsets, which only altered results for interpersonal measures.

#19

page 30, line 520: What estimates did the authors use for these analyses? The estimates shown in Table 2 or the "simple effects" that went into the analyses that led to Figure 2? I assume the former values were used, but please clarify this. Also, if my assumption is correct, then as far as I can tell, listwise deletion (due to incomplete information about the predictor variables) led to the removal of 120 - 45 = 75 estimates for T1 and 95 - 41 = 54 estimates for T2. Is that correct? If so, then this should be mentioned as a limitation.

Answer

The analyses were based on the ostracism effect across all 120 studies (as in Table 2 column d T1). However, due to listwise deletion the number of effects indeed reduced the number of effects included and now reads:

To inspect for structural and sampling effects of the studies, we ran mixed-effect models on the 120 ostracism effects, on both the first and the last measure. Due to listwise deletion, only 45 of 120 effect sizes remained on the first measure and 41 of 95 effect sizes for the last measure.

#20

pages 30, line 527: The dfs for the Q_E-test are 32. With k = 45, this implies that the model must have contained 45 - 32 = 13 fixed effects (including the intercept). However, in Table 4, I only count 12 coefficients.

Answer

We thank the reviewer for noting this error. The dfs should indeed be 33. This is now adjusted in the revised manuscript.

#21

page 31, line 537: The dfs for the Q_M-test are 12. Assuming that the intercept was not part of the coefficients tested, this implies that the model included 13 fixed effects. However, I only count 12 coefficients in Table 5.

Answer:

We again thank the reviewer for noting this error. The df should be 11 and is adjusted in the revised manuscript.

#22

page 31: Please report the results from the Q_E-test here as well.

Answer:

We added the results. On page 32 of the revised manuscript we now say:

Q_E (29) = 214.69, p < .0001

#23

Tables 4 and 5: For a categorical predictor with more than 2 levels, please provide a test of the factor as a whole (i.e., an omnibus test of the coefficients corresponding to the factor). Also, the tables only show the results of tests comparing each level against the reference level, but there may be significant differences when comparing other levels against each other. Please examine/report this.

Answer:

The Q_M test is an omnibus test and is reported. The dummies are indeed only compared to the reference group. Moreover, we already included confidence intervals in the original version of our manuscript. These CIs indicate that all comparisons between these dummies will yield similar results (overlapping CIs). Indeed, the requested analyses confirmed this:

If we only look at the countries, QM(df = 2) = 0.3494, p-val = 0.8397, first measure, QM(df = 2) = 2.6394, p-val = 0.2672, last measure.

If we only look at the different needs scales, QM(df = 4) = 6.0702, p-val = 0.1940, first measure, QM(df = 4) = 0.4257, p-val = 0.9803, last measure.

Because these analyses provide the same information as the overlapping confidence intervals we decided not to incorporate them in the revised manuscript.

#24

page 41, line 738: I don't understand what the authors mean by "difference index" or how this was coded. What "value" are the authors referring to when they write: "coded value on first measure minus coded value on last measure"? In fact, I have a hard time understanding this entire paragraph.

Answer

We thank the reviewer for this comment. We wanted to explain that differences in findings between first and last measurement could not be attributed to differences in types of dependent variables. We now write (on page 41-42):

Importantly, we did observe that the confidence intervals of both the first and last measure did not overlap, suggesting that there is a difference in effect size between first and last measure. The question then is whether this difference is indeed caused by time of measurement or in part caused by the type of measurement used across the two different time points. This explanation can be addressed by inspecting whether the composition of measures is different across time points. On the first measure 0.84 was intrapersonal self-report, 0.02 was intrapersonal physiological, 0.01 was intrapersonal other, 0.08 was interpersonal anti-social, 0.03 was interpersonal pro-social, and 0.01 interpersonal other. On the last measure 0.79 was intrapersonal self-report, 0.04 was intrapersonal physiological, 0.02 was intrapersonal other, 0.05 was interpersonal anti-social, 0.08 was interpersonal pro-social, and 0.01 was interpersonal other. This shows that the different types of dependent variables are similarly distributed across time points (maximum discrepancy of 4.9 percentage points). Substantive differences in proportions of measures across time points are minimal and thus form an unlikely driving force for our findings.

Minor Issues:

#25

Maybe this term is well understood by the intended target audience, but I find the term "cross-cutting variable" less than clear. Why not just call them "other factors" or something along those lines?

Answer

The term cross-cutting factor is a standard term in the Cyberball field. It refers to design in which the ostracism manipulation (inclusion vs ostracism) is orthogonally crossed with another manipulation (e.g., ingroup vs outgroup). Additionally, because we also include other moderator variables (i.e., time, structural, sampling), we use "cross-cutting" as a term to prevent confusion. Cross-cutting refers to the 52 studies that explicitly manipulated a factor in the experimental design. The other moderator variables (e.g, time, structural, sampling) were investigated for all 120 studies.

#26

page 3, line 47: The "(4)" is superfluous (or also number the other moderator types).

Answer

Adjusted

#27

Page 11

page 3, line 53: Write out "i.e." when used outside of parentheses.

Answer

Adjusted (also checked rest of i.e. occurrences)

#28

page 3, line 54: "an unique" should be "a unique" (the use of "a/an" is not based on the spelling of the first letter of the following word, but its pronunciation).

Answer

Adjusted

#29

page 7, line 150: "set-up" should be "set up" (set-up or setup is a noun).

Answer

Adjusted

#30

page 9, line 182: "extend" should be "extent" (the latter is the noun). And the more common phrasing would be "to a large extent".

Answer

Adjusted

#31

page 11, line 226: Write out the acronym (SPSP) the first time it is used.

Answer

Adjusted

#32

page 13, lines 291 and 293: Since you are giving examples here ("e.g.,"), the "etc." at the end is superfluous.

Answer

Adjusted

#33

page 14, line 301: Missing comma after "e.g.".

Answer

Adjusted (checked all occurences of e.g.)

#34

Table 1, table notes: I think the "whereas column wise" should be "whereas row wise".

Answer

Adjusted

#35

page 41, line 754: "conditional on that these measures are valid" is very odd phrasing.

Answer

Deleted this sentence.

#36

The Oxford comma is used inconsistently throughout the manuscript.

Answer

We checked the manuscript for consistency and adjusted where needed.

Appendix:

#37

1) df_w needs to be defined.

Answer

Adjusted. Added that this is equal to conditions minus 1.

#38

2) The top part of a fraction is called "numerator", not "nominator".

Answer

Adjusted

#39

3) Isn't the first term in the numerator the ostracism effect *in the non-moderated/control condition* (and the second term is the effect in the moderated condition)?

Answer

We calculated it in the order we describe. It can also be done the other way around, which would lead to a change in interpretation but equal results.

Page 13

#40

4) In what sense does Delta-d "correspond" to partial eta-squared of the interaction? Numerically it cannot be the same (partial eta-squared must be between 0 and 1, while Delta-d as defined is not a proportion and may be larger than 1 and can be negative).

Answer

When the resulting d is transformed into a squared correlation coefficient it gives the exact same value. This is highlighted in the Appendix and now reads

When transformed to a squared correlation coefficient, this Δd corresponds to the partial eta-squared of the interaction.

#41

5) Please add ^2 to s_g and s_d to make it clearer that these are variances.

Answer

Done.

#42

Final comment: In the spirit of open science, I appreciate the use of OSF and the authors' transparency in conducting this meta-analysis.

Answer

Thank you. We also like to thank the reviewer for the thorough review and thus for making this a better manuscript.

Reviewer #2:

#1

Overall this study looks competently executed and acceptable for publication. My only real concern is that authors could have done more to explore and account for the variability in their data. The meta-analysis demonstrates that the variability was considerable, but beyond establishing that moderators exist, the researchers appear to be not overly concerned with the question what is causing this variation. That leaves me slightly unsatisfied at the end: all this effort to conduct a meta-analysis, and the main thing we learn is that (a) the effect of rejection is strong (something we knew because it has been shown time and again), (b) the first sharp shock diminishes over time (new to me, but then I'm not an expert), and (c) the intensity of that shock depends… If authors were willing to stick their finger out a bit more and clarify just what this depends on, I'm sure I would find the study more valuable than it is now. I don't care if their hypotheses were deposited beforehand: exploring is a scientists' duty, as much as hypothesizing in advance (e.g., Tukey). But to be clear: this is just meant an encouragement; it's very much up to to the authors to decide what course of action to pursue.

Answer

We thank the reviewer for his/her kind words and regarding the manuscript as competent and acceptable for publication. We agree with Reviewer #2 that exploring the data is a valuable avenue for any study, including this meta-analysis. As a matter of fact, we were also puzzled by the heterogeneity in the data and we therefore conducted several exploratory analyses to understand this heterogeneity. The most important exploratory analysis that we conducted was the one in which we selected the most homogenous subset possible (i.e., only immediate fundamental need measures, 30 throws, 3 players), but still found high heterogeneity. Meta-regressions also failed to indicate any explanation for the heterogeneity. We agree that further exploration is definitely interesting, but also believe that we exhausted all possibilities that were available to us in the current dataset.

Some other points that would help authors improve the paper up to a level that would match my expectations for PLOS One standard mainly concern the quality of the writing and the care about the argument being made. The introduction reveals that authors could have spent some more care writing (and perhaps thinking about) their subject. Suffice to say that it's important to be precise. Some examples:

#2

"Cyberball participants simply do not obtain a ball and thus need to infer that they are excluded" I think authors are trying to say something about implicit and explicit exclusion here. I also think they are trying to say something about acting together versus communicating with each other. But it's not being said.

Answer

This sentence was deleted, because the preceding sentence already contains the information.

#3

The sentence “This focus on ostracism makes it an unique paradigm...” is clearly erroneous, because it is not the focus on ostracism that makes cyberball unique.

Answer

The first paragraph in the Historical background section is changed into:

Cyberball was introduced in 2000 as a means to study ostracism, that is: being excluded and ignored [1]. This focus of Cyberball on ostracism sets it apart from other paradigms that are tailored to study rejection, such as the future life rejection [2], the get-acquainted paradigm [3], and the autobiographical memory manipulation (i.e., remember a time when you were excluded [4]). The difference is that participants in Cyberball are not explicitly informed that they are excluded whereas in the other paradigms participants are provided a reason pertaining to why they are excluded.

#4

Further on, a sentence such as “research suggests that most people are ignored and excluded at least once a day” sits happily side by side with the sentence “research on school shootings has suggested a direct link between ostracism and revenge”. This could be spelled out more clearly. If everyone is a victim of exclusion, then obviously those who go on a shooting spree are, too. So is the point that ostracism is a frequently occurring post-hoc justificationfor this kind of behavior?

Answer

We adjusted the sentence. It now reads:

The social relevance is further evident in that ostracism not only affects the person who is ostracized (intrapersonal effects), but often also others (interpersonal effects). As a grim example, research on school shootings has suggested a direct link between ostracism and revenge. People who were ostracized may retaliate by murdering those responsible and sometimes even innocent bystanders [5].

#5

Further on authors write “This initial response is theorized to be socially painful, threatening [9] and easily detectable due to evolutionary over-sensitivity to cues of ostracism [12].” In a sentence such as this, please carefully distinguish phenomenon and hypothesis. There is abundant evidence for the first inference, but the evolutionary origins of this phenomenon can only be inferred indirectly from its existence and prevalence.

Answer

We adjusted the sentence. It now reads:

This initial response is theorized to be socially painful, threatening [9] and, following overdetection theory [12], should be easily detectable due to evolutionary over-sensitivity to cues of ostracism.

#6

It is stated that all selections and hypotheses were preregistered on OSF. But what is not spelled out is whether authors tried to learn something new from their data by exploring it?

Answer

We explored several avenues. For example see reviewer #2, answer #1, but also answer below (reviewer #2, answer #7)

#7

"Examples of interpersonal measures are donations to charity, helping behavior, money allocations in economic games, and aggression measures such as irritating sounds blasts or hot sauce allocation." Please split the effects of positive and negative behaviors—they are qualitatively too distinct to be lumped together in this way. Later on I noted that K=10 for these studies (?). If small K was the reason for lumping things together please explain the criteria and total K in this section to help readers understand your decision making process.

Answer

These were indeed split into positive (pro-social) and negative (anti-social) behaviors initially and were indeed lumped together due to small K, hence, low power for detecting moderation effects. For the first measure, there were 14 interpersonal measures, of which 4 are positive and 10 negative. For the last measure, there were 14 interpersonal measures, of which 8 are positive and 6 negative. We added a sentence in the manuscript to clarify this. One page 8 of the revised manuscript we now say:

These were initially coded into pro- and anti-social, but were collated into the category interpersonal due to small *k* the first measure (4 and 10, respectively) and last measure (8 and 6, respectively).

#8

For various decisions to include or exclude studies or factors, please provide an indication of the number of studies affected by your decision. E.g., "continuous variables that were dichotomized into factorial levels were also collapsed due to the many problems dichotomization can cause". How many studies were collapsed in this way? I'm trying to assess the impact of your coding decisions.

Answer

This collapsing occurred a total of four times, for the studies from (i) Stock 2011, (ii) two studies from Boyes 2009, and (iii) Zadro 2006. We added this number in the manuscript on page 10..

Some other minor points:

Page 17

#9

“we used the metafor package”: include version.

Answer

Version 1.9-5. Added in the manuscript.

#10

I do not understand this sentence: “Model indicates that the first measure was indeed reflexive and the last measure reflective.”

Answer

The model pertained to a subset included throughout the analyses. To avoid confusion we rewrote the note under table 3 to read similar to Figure 2

The subset labeled “All” contains all measures. The subset labeled “Fundamental” contains only fundamental need measures. The subset labeled “Intrapersonal” contains all intrapersonal measures. The subset labeled “Interpersonal” contains all interpersonal measures. The subset labeled “Model” contains those where first measures is immediate and last measure is delayed. See Supplement S4.

#11

“meta-analyses” is plural

Answer

Adjusted

#12

“by a large extend”
= to a large extent

Answer

Adjusted

Reviewer #3: This study is a system review and meta-analysis of cyberball studies for effect size of ostracism. The manuscript is well-written and provides many detailed information for readers. The statistical analysis is rigorous and well-thought. The primary and secondary hypotheses are clearly stated. The results and discussion are also clearly presented. I have following comments.

We thank the reviewer for his kind words and stating that our analyses are rigorous and the manuscript is well-written.

#1

1. First, I appreciate the authors' efforts in providing detailed information about the data and implementation, which greatly improve the transparency and reproducibility of the research. More importantly, the information is very helpful for readers to have an objective view of this study.

Answer

Thank you for your kind words.

#2

2. I would suggest moving the "code procedure" sub-section in Method section to supplementary. Although the code procedure is very important and helpful for some readers, it is too technical for most of readers.

Answer

Although we understand the concerns for the technicalities, the supplement is meant for additional information only, while we consider the coding a crucial aspect of our method. We had thorough discussions on whether it was possible to have directional coding in spite of the bidirectionality of the expected effects and we think a reader will want to know how we were able to make directional claims despite this variety of measures and predictions. Hence, we think it is vital to retain this in the main manuscript.

#3

3. I suggest adding a figure for study inclusion criteria. Many system review and meta-analysis paper in PLoS ONE use a figure to demonstrate the procedure for selecting studies.

Answer

The manuscript contains the PRISMA flowchart in the supplemental materials that addresses this point. We added the flowchart in the manuscript.

#4

4. It's better to present the information in Table 2 as a forest plot, while putting the table 2 in supplementary. A forest plot summarizes the information and gives readers a intuitive understanding.

Answer

Page 19

We agree that a forest plot gives an intuitive overview of the effects. However, we think that the forest plot across 120 effects will be too sizable. More importantly, the American Psychological Association prescribes that meta-analyses are to report the data on which main analyses are performed in a table. We therefore think it is more informative to retain the current format.