How strong is our evidence? Evidential value and publication bias in research on social media use and self-esteem.

Amber van Anen

Snr 2071056

Master's Thesis

Communication and Information Sciences

Specialization Communication and Cognition

Department Communication and Cognition

School of Humanities and Digital Sciences

Tilburg University, Tilburg

Supervisor: Dr. L. Coenen

Second Reader: Dr. M. Pollmann

8 June 2022

Abstract

Social media research has especially contradicting findings. An important line of research within social media research is the relationship between self-esteem and social media use. Research has reported both a negative and positive effect of social media use on self-esteem as well as an impact of self-esteem on social media use. These contradicting findings are a reason for concern, especially keeping publication bias and p-hacking into account. Publication bias and p-hacking are very prominent social sciences, and replication research has shown that many results are not significant during these attempts. Communication science is relatively new to replication, so it is important to investigate this as well. Three methods were used to test this problem: a systematic literature review and both a p- and z-curve analysis. While the systematic literature review highlighted a lack of transparency, mostly in reporting scales and statistics, the p- and z-curve indicated sufficient evidential value and the likely absence of heavy publication bias. This thesis will furthermore discuss the implications of the evidential value and methodological weaknesses. For the OSF registration, see https://osf.io/r3a7c/

Keywords: Social media use, self-esteem, evidential value, p-curve, z-curve, publication bias, p-hacking

How strong is our evidence? Evidential value and publication bias in research on social media use and self-esteem.

Social media are an essential part of daily life. With only Instagram having a daily count of 500 million active users, these social media platforms have a significant role in our society (Dean, 2022). Even though these platforms were designed to entertain and keep social relationships (Instagram, n.d.), the negative consequences of social media usage quickly became a concern among users, especially regarding mental health. For instance, a BBC article summarizing the harmful effects of social media links it to various topics like stress, depression, and lower self-esteem (Brown, 2018). Other concerns cover the addictive aspects of social media and (negative) social comparison (Williams, 2021). This negative impact seems to be confirmed by much other media coverage over the last years (Walton, 2021; Headspace, n.d.).

The concern with the effects of social media on mental health is also visible in the available scientific literature. One particularly popular topic is social media usage's negative effect on self-esteem (Berry et al., 2018; Romero-Rodriguez et al., 2020; Woods & Scott, 2016). For example, spending more time on Facebook leads to lower levels of self-esteem (Jan et al., 2017). These findings may sound worrying, especially because self-esteem is known to be related to depression and anxiety, while lower levels of self-esteem are related to unhappiness and lower self-worth (Holloway et al., 2016).

However, just because there seems to be a decent amount of evidence in the scientific literature does not mean this is adequate evidence of this relationship. Once a paper gets published with significant results, the community will most of the time see this as proper evidence of this effect, but this is known to be a false assumption potentially; significant findings in the literature need not reflect an actual effect but can also signal a false-positive effect (Ioannidis, 2005). One reason for this false-positive finding is the error rate, the 5%

chance that a significant result will be found if the null hypothesis is true. This error rate is mainly accepted as it is a slight chance of finding an incorrect result, but it is something to be aware of (Ioannidis, 2005).

Another possible concern surrounding evidential value in research is publication bias. This bias is the tendency of scientific journals to be more likely to post papers with significant findings than papers that did not find significant results (Dickersin, 1990). Non-significant results are generally not seen as a relevant addition to the scientific field (Sterling et al., 1995). Because of this publication bias, researchers might be more likely to manipulate their findings to be able to get published; this is called p-hacking (Head et al., 2015). Manipulated, or p-hacked, results can be hard to recognize and, of course, replicate because of their false information, like incorrect effect sizes and an increase in the error rate (Francis, 2012). The existence of publication bias has been documented in many fields, such as psychology and the medical field, and has been a driving force behind the replication crisis (Freese & Peterson, 2017). In psychology, there were doubts about previous findings, and as a result, mass replications were conducted (Maxwell et al., 2015). Many results were found not to be replicable (Foster & Deardorff, 2017).

In communication science, the replication crisis has received comparably little attention. However, because of the overlap of methods between social sciences, it can be assumed that communication science suffers from the same problems in replicability as psychology (Dienlin et al., 2020). In the literature on social media and self-esteem, for instance, a few meta-analyses were conducted, but none ever focused on publication bias and evidential value specifically (Huang, 2021; Saiphoo et al., 2020). This is surprising, especially because there seems to be some inconsistency in the available literature. Even though there is a consensus that social media leads to lower self-esteem (Jan et al., 2017; Romero-Rodriguez et al., 2020; Woods & Scott, 2016), there is also literature proving that self-esteem is a predictor of (problematic) social media use (Andreassen et al., 2017; Schivinski et al., 2020). It has even been suggested that social media use has a positive effect on self-esteem, even though there is less literature surrounding this topic (Shaw & Gant, 2002). That is why this paper is trying to investigate how strong the evidential value is in the literature surrounding the relationship between social media use and self-esteem, and if there is any indication for publication bias.

Theoretical Framework

The relationship between social media on self-esteem

Historically, the literature on media effects has always emphasized the potentially negative, rather than positive, consequences of media use. For instance, from the 1960s onward, the literature on television effects has often portrayed television as an instrument making children violent in the 1990s; later, the same discussion was held with videogames (Anderson & Warburton, 2012; Savage & Yancey, 2008). Over the past decade, research has shifted its attention to social media; here, too, the potentially negative consequences have been at the center of attention. Especially the effects on self-esteem have been discussed at length: this makes sense theoretically, as exposure to social media has proven to elicit social comparison and unrealistic ideals (Alfasi, 2019; Moningka & Eminiar, 2019; Vall-Roque et al., 2021; Woods & Scott, 2016). Despite this intuitive theoretical rationale, the findings on the matter appear to be mixed – at best.

The dominant narrative in published studies seems to be that there are, in fact, negative effects. Social media usage has been found to cause lower levels of self-esteem (Krause et al., 2019; Saiphoo et al., 2020). Self-esteem is revised through three processes; social comparison, social feedback, and self-reflection (Krause et al., 2019). Since social media facilitates all of these processes with users comparing themselves with other users, getting feedback in the form of likes, and reflecting on their content, it is logical that the effects on self-esteem are visible. Considerable research also focuses on these processes and their negative impact on self-esteem, with social comparison seemingly the most prominent process (Krause et al., 2019). For example, the study by Alfase (2019) has shown that browsing Facebook causes upwards social comparison and, as a result, lower state self-esteem. These findings are in line with similar research on this topic (Lui et al., 2017; Ozimek & Bierhoff, 2019; Liu et al., 2017; Schmuck et al., 2019). Not all research focuses on specific underlying processes, but a direct link between SNS use and self-esteem has also been found (Jan et al., 2017; Malik & Khan, 2015; Romero-Rodriguez et al., 2020; Stapleton et al., 2017; Woods & Scott, 2016).

In contrast to this dominant narrative, several studies have found the sign of effects to be positive rather than negative**.** A notable example of this is the study of Wilcox & Stephen (2013), who demonstrated in multiple studies that using social media can actually enhance self-esteem (Wilcox & Stephen, 2013). This finding is not unique; the social support one can get from social media can, for example, lead to an increase in self-esteem (Shaw & Gant, 2002). Similarly, positive online feedback can positively affect self-esteem (Burrow & Rainone, 2017).

Even other studies have suggested that the causal order should be inverted. For example, Andreassen, Pallesen, and Griffiths (2017) looked at problematic social media use predictors. This study discovered that people with higher levels of problematic social media use also had lower self-esteem (Andreassen et al., 2017). Problematic social media use is described as excessive use of social media platforms or smartphone addiction (Andreassen et al., 2017; Romero-Rodriguez et al., 2020). They saw this relationship as a result of people with lower self-esteem wanting to feel more self-worth by getting more likes, for example, and thus using more social media (Andreassen et al., 2017). Other research that examined different predictors for social media addiction also found low self-esteem as a significant predictor (Hawi & Samaha, 2018).

This pattern of inconsistent findings (and inverted causal theories) is exemplary for research on media effects in general. The same irregular patterns were found in television research & video game research: where the initial primary focus was put on the negative effects, later research even showed positive effects of these media, like an increase in pro-social behavior (Anderson & Warburton, 2012; Mares & Woodard, 2005). The question this raises is twofold. First, why does the media effects literature embrace all these (seemingly inconsistent) patterns? And second, how convincing is the existing evidence, then? Typically, the literature addresses these two questions by emphasizing the complexity of media effects. For instance, the Differential Susceptibility Model proposes that all media effects are conditional; patterns may occur depending on dispositional, developmental, and social variables (Valkenburg & Peter, 2013). So, the selection of and responsiveness to media can vary under different personal and social conditions (Valkenburg & Peter, 2013). This model has also been applied to social media and self-esteem: Cingel, Karter, and Krause (2022) noted that many of the moderators used in this line of research are dispositional susceptibility factors, like age, gender, or need for popularity. So, this model states the effects of social media on self-esteem might be conditional.

Others have been more critical, suggesting that the inconsistent patterns are caused by the media effects literature being methodologically and conceptually weak. For instance, Orben and Przybylski (2019) recently showed that the effects found in research on the impact of social media on mental health are small and unconvincing. This is argued to be the result of methodical flaws surrounding this research. One potential flaw is the statistical freedom within data analysis where researchers are allowed to make many decisions during the analysis, with only the final statistical pathway being shown in the paper (Orben &

Przybylski, 2019). Another flaw within social media research is that in large-scale datasets, with, for instance, self-reports, a slight variation between survey items can lead to significant results, indicating convincing evidence (Orben & Przybylski, 2019). Also, cross-sectional research, which is used a lot within behavioral science, relies on correlational evidence, but this method of research is frequently applied causal relationships (Orben & Przybylski, 2019). However, despite these criticisms, there have been few systematic attempts to map the severity of conceptual and methodological problems. This thesis will therefore attempt to address this gap by (1) providing a conceptual analysis of the relationship between social media and self-esteem and (2) reviewing the (lack of) statistical evidence in terms of p- and z-curves.

Conceptual issues

Because of this internal inconsistency in the literature surrounding the relationship between social media use and self-esteem, it is interesting to analyze the potential conceptual weaknesses. As discussed before, the replication crisis seems to indicate that analyzing previous research can stress potential fallacies (Dienlin et al., 2020). Freese and Peterson (2017) looked at the replication crisis and the implications this might have on social science. They came up with four types of replication: verifiability, robustness, repeatability, and generalizability (Freese & Peterson, 2017). Verifiability in this context is a way of inspecting the results of a paper and analyzing whether these results are proper results by looking the data. Robustness relates as the authors state, to examining the data again and analyzing whether the result is an actual effect or just the result of statistical steps taken by the researchers (Freese & Peterson, 2017). Verifiability and robustness can be studied with old data, whilst repeatability and generalizability have to be inspected with new data. Repeatability and generalizability study the original findings by either using the same steps as the initial paper or testing if the findings can be repeated in a different context or method (Freese & Peterson, 2017).

Because of the meta-analytic nature of this study, the types of replication that are most relevant are verifiability and robustness. Verifiability can relate to transparency like publishing the data or practicing open science. If present, one can check if the data logically leads to the results shown in the paper. To ensure transparency within verifiability, one must ensure that the article's findings are actual findings and not just a result of p-hacking. Analytical steps and decisions need to be clear like clear conceptualizations and a transparent results section (Freese & Peterson, 2017). These concepts are similar, with both of them surrounding the topic of replication. However, verifiability refers more so to the transparency of the entire paper, while robustness focuses on the analyses. With a focus on verifiability and robustness, one can check the conceptual transparency of previous studies. For this reason, the current study will answer the following sub-question:

To what extent is the literature surrounding the relationship between social media use and self-esteem conceptually transparent, based on the concepts of robustness and verifiability?

Statistical issues: publication bias and p-hacking

In theory discussed above and in general, conclusions are based on scientific evidence. However, when can a result be labeled as evidence? In most scientific disciplines, significance is tested with a p-value as a result, with a p-value lower than 0.05 indicating significance (Ioannidis, 2005). P-values are the basis of null hypothesis significance testing. The null hypothesis states that there is no effect, where the p-value is the chance that an effect is found while the null hypothesis is true; this is related to the error rate since this is the chance a significant effect is found while there is no effect. (Head et al., 2015). So, if a p-value is lower than 0.05, it will get interpreted as a difference in groups or as an effect, while a p-value higher than 0.05 supposedly means that there is no difference or effect.

However, there is considerable criticism surrounding this way of testing for significance. Ioannidis (2005) criticized the p-value in a paper discussing that most published findings are false. In this paper statistical simulations were discussed, calculating the chance of finding true or false effects based on, e.g., type I and type II errors. These simulations showed that the chance of a false positive (discovering a significant result while there is no effect) is high for many various study designs and contexts, depending on, for instance, pre-study odds, sample size, and the flexibility of designs (Ioannidis, 2005). Amhrein, Greenland, and Mcshane (2019) even proposed for statistical significance to be abolished in total. They argue that the existing way of thinking about statistical significance makes researchers more likely to cause false interpretations because of thinking in either effects or no effects; thus, they suggest quitting categorizing (Amhrein et al., 2019). However, the biggest issue surrounding the p-value might be publication bias; this is the tendency of papers with significant results to be published more than papers without significant results (Dickerson, 1990).

As a result of publication bias, researchers might be more likely to influence their data to get a significant p-value; this is known as p-hacking (Head et al., 2015). P-hacking is the misreporting or influencing of data to get a significant result (Head et al., 2015). Head et al. (2015) used text mining to look at published papers from 14 different disciplines and found evidence that p-hacking frequently happens in all fields (Head et al., 2015). P-hacking causes the error rate of the p-value to go up drastically (Ioannidis, 2005). Since p-hacking can cause false-positive results (Head et al., 2015; Ioannidis, 2005), this is a big concern for the evidential value of science.

Therefore, it is necessary to understand how p-hacking can be identified. Wicherts et al. (2016) looked at p-hacking and recognized it in different stages of the scientific process. They called all of these various choices that one could make researcher degrees of freedom

(DF) (Wicherts et al., 2016). In total, there are 34 DF in five research stages. In the hypothesizing phase, one can conduct exploratory research and later present this as confirmatory research, also known as HARKing, hypothesizing after results are known (Wicherts et al., 2016). In the design phase, researchers can measure variables in different ways or measure additional variables that one can later use as primary outcomes (Wicherts et al., 2016). In the collection phase, the DF's concern insufficient blinding and randomization. Most DFs exist in the analysis phase. As discussed before, variables can be measured and used in different ways to get a lower p-value. Also, one can decide how to deal with outliers or assumption violations in an ad hoc manner, depending on the desired outcome (Wicherts et al., 2016). In the reporting phase, the last DF's contain choices in replication and misrepresenting the study (Wicherts et al., 2016).

Simmons, Nelson, and Simohnson (2011) also tested these DF's. They analyzed four different degrees of freedom and their influence on the p-value: measuring a dependent variable in two ways, deciding the sample size based on when the effect was significant enough, controlling for covariates, and dropping conditions (Simmons et al., 2011). Practicing these DFs led to an increase of false-positive effects between 11.7% and 50%. Applying all of them leads to a false positive rate of 67% (Simmons et al., 2011).

It is essential to be aware of these DFs to recognize p-hacking. However, many p-hacking techniques are not recognizable when reading an article. For example, HARKing can only be recognized if the original plan is open to the public (Wicherts et al., 2016). A solution to this problem might be open science. Open science can be practiced by pre-registering research so the public can see early versions and the original research plan (Foster & Deardorff, 2017). Another important concept of open science is the availability of research data. Practicing science in this manner means more transparency, and p-hacking might be less prominent.

Open science and replication might be reasonable solutions, but open science is only a solution for future research, and replication is very time-consuming. Another way of checking the evidence for p-hacking is through statistical analyses. Simmonson, Nelson, and Simmons (2014) created such an analysis based on the distribution of the p-value – the so-called p-curve analysis. A p-curve analysis compares the distribution of p-values in the literature on a specific topic to the distribution of p-values an actual effect would have. Thus, by conducting a p-curve analysis, one can check to see if there are more significant findings than there should be, indicating p-hacking and publication bias (Simonsohn et al., 2014).

Based on the prominent existence of p-hacking and publication bias, the question can be asked if these concepts are present in the literature surrounding the relationship between self-esteem and social media use. Previous meta-analyses have also tried to check this literature for publication bias (Saiphoo et al., 2020) but did so using the Trim and Fill method; this method focuses on effect size instead of significance as it assumes the selective reporting of papers is based on either small or big effect sizes, even though it is proven that the basis of publication bias is the statistical significance (Simonsohn et al., 2014). They found no indication of publication bias; this can result from the criticized choice of method. Because of this, it is still important to check for publication bias and p-hacking using a p-curve analysis. That is why the second sub-question is:

To what extent is p-hacking and publication bias present in the literature surrounding the relationship between social media use and self-esteem, based on a p- and z-curve analysis?

Methods

Systematic literature review

Search strategy

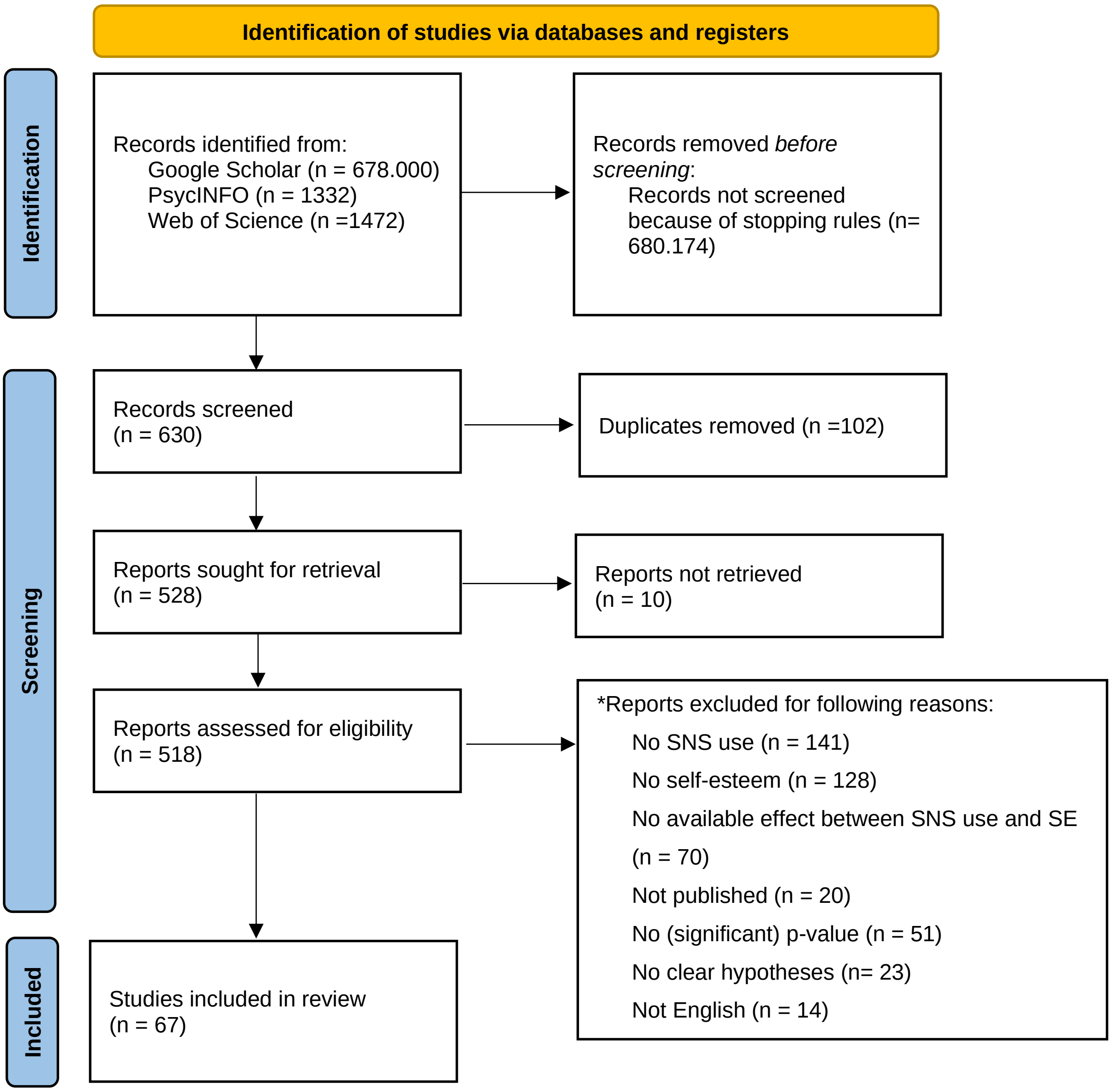
To sample relevant literature on the relationship between social media use and self-esteem the following databases were used: Psychinfo, Web of Science, and Google Scholar.

The search terms were "social media" OR "social network*" OR SNS OR Facebook OR Twitter OR Instagram OR Tumblr OR Pinterest OR Flickr AND "self esteem." Similar research on social media effects also used these terms (Saiphoo et al., 2020). All abstracts were screened to check for relevancy; if they were deemed irrelevant, the article was not included. Relevancy in this context meant that studies needed to measure self-esteem *and* social media and assess a relationship between the two. If the article came through the (manual) screening, it was added to a Microsoft Excel database where all relevant article information was summarized and stored. Given availability and time constraints, not all articles returned by the search engines could be included in the research. The search process was therefore stopped if, for all three search databases, no more relevant articles came up or only doubles were found, which was after results 330 on Google Scholar, 200 for Web of Science on 100 on PsychInfo. Since three different search databases were used and the screening stopped only at doubles or irrelevant articles the search process presumably resulted in a relatively representative database. The details of the search strategy can be found in Figure 1, showcasing the steps of data collection through the PRISMA model; a model for reporting systematic reviews and meta-analyses (Liberati et al., 2009).

After screening the articles for verifiability and robustness, additional information was extracted to prepare them for the p- and z-curve analyses. First, Doubles were found and excluded. Second, the articles were assigned to separate categories, depending on the research focus: Self-esteem as an independent variable, self-esteem as a dependent variable (negative effect), and self-esteem as a dependent variable (positive effect). As extra relevant information, the type of SNS use was also coded into either 'general SNS use', 'problematic SNS use', or 'specific SNS behavior'.

Figure 1

PRISMA model of data collection



**Four articles are removed during the analysis since the results were irrelevant for the analysis; they did not meet criteria three (n=2) and six (n=2) after further inspection.*

Inclusion criteria

For an article to be included in the analysis, it needs to check the following criteria.

1. Some form of social media use needed to be measured. Three different categories were discovered. The first one is regular SNS use; this can be as vague as ‘using social media’ or more specific to a platform like Facebook usage, like “Facebook intensity”. The second category is problematic SNS use; this can be addictive social media use or

(negative) social comparison. The last category is specific SNS behavior, like taking selfies or posting certain content. It was not included if social media use does not fall under these categories, for instance, cyberbullying or the intention to quit social media.

2. Self-esteem needed to be measured and used in the research. In the context of this study self-esteem is an individual's evaluation of their overall value and worth.
3. Studies needed to test a relationship between social media use and self-esteem. This effect could be a total effect or a specific path in a mediation model, but an effect from one to the other was required.
4. Since the purpose of the p- and z-curve analysis is to test for publication bias, the article needed to be published. Unpublished papers and theses were not included.
5. A p-value, or information sufficient to reconstruct the p-value, needed to be available.
6. A clear hypothesis about the relationship between social media use and self-esteem needed to be tested. This means that exploratory analyses or hypotheses without a direction (e.g., a correlation in a correlation table) were excluded.
7. Only articles published in the English language were included.

Replicability Analysis

Many aspects influence the replicability of an article. It is impossible to check the literature for all of these aspects, as some are not even recognizable. For this reason, the analyses reported here focus on two main criteria for replicability, as highlighted by Freese and Peterson (2017). These criteria were verifiability and robustness. The first category, verifiability, refers to research transparency and data availability. Verifiability was studied by checking (1) if papers were available in open science databases, and (2) if the data were made available, and (3) if the study was preregistered. The second category was robustness, with clear conceptualizations and transparency in the data analysis being the main focus (Freese &

Peterson, 2017). To study potential robustness fallacies, missing information or ambiguities in the methods and results section of the studies were noted and summarized in the systematic literature review. A second coder also coded a small portion of the literature (n=7) to improve the coding validity. Overall, the second coding seemed similar based on face validity. As the coding was a more interpretative process of noting any remarkable findings based on verifiability and robustness, no metrics could be calculated. However, similar patterns were discovered as in the original coding, like the missing statistics and referring to tables, and no big differences in coding were observed.

It is essential to note that any ambiguities or 'flaws' in this review are not to accuse the corresponding researchers. These articles are used as examples of patterns in research that are not entirely transparent or highlight possible methodological weaknesses, but the discussed research is not problematic.

P & Z-curve analysis

P-curve analysis

The p-curve analysis is designed by Simonsohn, Nelson, and Simmons (2014). This analysis uses an empirical distribution of significant p-values and compares it to the distribution of p-values under the null hypothesis (Simonsohn et al., 2014). The logic behind it is that the distribution of p-values tells a lot about the type of effect; true effects lead to right-skewed p-curves, no effect leads to flat p-curves, and p-hacked results lead to left-skewed p-curves (Simonsohn et al., 2014).

P-curve analyses do not rely on significant p-values from published papers per se. Instead, they rely on the *test statistics* associated with those significant p-values, which are then used to recalculate the p-values. This approach is useful as it bypasses the problem of p-values often being reported imprecisely (e.g., by using asterisks to indicate that a result is significant at some significance level). Hence, test statistics were coded into the database, together with their associated hypothesis. Since some articles failed to report the test statistics or other relevant information, a few statistics were manually calculated before the analysis. For instance, in some articles, the degrees of freedom were not reported. When this occurred degrees of freedom were calculated based on the sample sizes and the type of statistic, and imputed in the analysis. Other articles lacked even more information, only reporting the p-value and the beta coefficient. In these instances, test statistics were calculated based on the estimated degrees of freedom, p-value, and sample size, using R-studio. The t-value was calculated based on the p-value, using *t* as the quantile function of the reported *p*.

As noted before, the selected articles were split into three groups: self-esteem as an independent and a dependent variable and positive effects on self-esteem. These groups were compared in the p-curve analysis to see if the research question at hand mattered to the existence of publication bias.

The p-curve analyses were conducted by using the p-curve app (http://www.p-curve.com/app4/). The app takes test statistics as input and produces a figure showing the distribution of p-values, together with binomial and continuous tests indicating if the literature has evidential value or not. The binomial test is a comparison of the observed portion of results that are $p<0.025$ to when there is no effect, and to an effect with 33% power. If the 33% power test is $p<0.05$, it can be concluded that the evidential value is inadequate or absent. The second tests, the continuous tests, use a method for first computing pp-values for each test and then converting Z-scores, known as Stouffer's Method, for a full (p-values<0.05) and a half p-curve (p-values <0.025) (Simonsohn et al., 2015). The half p-curve was added to check for ambitious p-hacking; p-hacking where researchers try to get a p-value even lower than 0.05 (Simonsohn et al., 2015). The scores for these curves are combined to draw conclusions about the evidential value. Comparing these scores can give an indication of evidential value; if the half p-curve is $p<0.05$ or both of the curves are $p<0.1$, evidential value is present. (Simonsohn et al., 2013). The p-curve disclosure table can be found in the Open Science Framework (https://osf.io/5fydc/).

Z-curve analysis

A z-curve analysis was also conducted to ensure to corroborate the robustness of the p-curve analyses. The z-curve analysis was created as a response to the p-curve and aimed to solve some of its shortcomings, such as p-curves tendency to overestimate power (Schimmack, 2021). The z-curve analysis was performed using the z-curve package in R (Bartoš & Schimmack, 2020). The analysis is quite similar, but it has advantages, such as extra estimations of the file drawer effect and a quantified measure of the strength of the evidence (Schimmack, 2021). Z-curve analysis uses z-scores, as the name indicates. These z-scores were calculated during the analysis, converted from exact p-values; using the quantile function. These exact p-values were retrieved from the output of the p-curve analysis, and manually entered in the z-curve analysis.

The z-curve analysis produces a few interesting results; the expected replicability rate, the false discovery risk, the observed discovery rate, and the expected discovery rate. The expected replicability rate indicates the percentage of studies that would give a similar result if replicated. The false discovery rate indicates the number of studies that possibly contain false positives instead of true effects. The last statistics, the observed and expected discovery rate, can be compared to draw a conclusion about the proportion of significant vs. non-significant findings (Schimmack, 2021). Note that the last two statistics are not relevant for the purpose of the current study: the primary analysis of this paper is the p-curve analysis, which uses only the distribution of significant results. Because of this, the proportion of (non)significant results reported by the z-curve is not representative of the actual data, meaning that the observed and expected discovery rates are not directly interpretable. The expected replicability rate and false discovery risk are nevertheless still appropriate.

Results

Systematic literature review

Data characteristics

There are 71 relevant articles in this study. All studies were published between 2011 and 2022. Of these articles, 11,9% were published between 2011 and 2013, 14,9% between 2014 and 2016, 41,8% between 2017 and 2019, and 31,4% between 2020 and 2022. Many articles only focus on a specific age group; mostly adolescents (n=5), teens (n=5), or college-age (n=13). Other articles have similar specific target groups, focusing only on women, problematic SNS users, or specific personality traits (e.g., narcissism or type D personality). Of all articles, 28 focus on self-esteem as a dependent variable, 24 as an independent variable, and 15 on the positive impact on self-esteem. Likewise, 26 articles measure SNS use as general SNS use, 33 as problematic SNS use, and 8 measure a specific SNS behavior.

Verifiability

Practicing open science is fundamental to the concept of verifiability. Preregistration offers the possibility of checking the initial research plan and all steps made before the end result; the paper. Within the literature of this paper, the number of preregistrations was very low; two out of the 71 papers (2,82%) were preregistered - the study of Valkenburg et al. (2021) and Shin et al. (2017). However, the registration of the latter study was unavailable.

Another essential concept in verifiability is public data. Public data allows for checking the results of papers and possibly replicating the findings. The number of papers in this research that published their data was higher than the number of preregistrations, ten out of the 71 papers (14,08%). Some articles offered their data in the appendix (O'dea & Campbell, 2011; Trifiro & Prena, 2021; Valkenburg et al., 2021; Wang et al., 2018). However, not all datasets were immediately available (8,45%). When datasets are made available, they appear to be mostly available only on request; six out of the ten studies only made their data available on request (Acar et al., 2020; Brandenberg et al., 2018; Lim et al., 2021; Koçak et al., 2021; Steinsbekk et al., 2021; Vall-Roque et al., 2021).

Robustness

As noted before, robustness in the context of this paper relies on (1) clear conceptualizations and (2) transparency in data analysis. Overall, the conceptualizations were clear and, in some manner, standardized; since all literature in this paper focuses on similar concepts, many of the conceptualizations and scales were the same. For example, self-esteem was primarily measured (87,32%) as a score on the Rosenberg Self-Esteem Scale, a scale with high reliability (Rosenberg, 1965). Other articles use scales for specific types of self-esteem (e.g. state self-esteem or appearance based self-esteem) or single item scales (Fox et al., 2021; Steinsbekk et al., 2021; Marengo et al., 2022).

The ambiguity of conceptualizations increases with measurements of SNS use. There are standardized scales for SNS use, like the Facebook Intensity scale (Ellison et al., 2007) (16,90%) or the Bergen Facebook Addiction Scale (Andreassen et al., 2012) (18,31%). Still, other researchers decided to create their own measurement instruments. In these measurements, there can be a lack of transparent reporting, for example, not noting how the questionnaire relates to the eventual score (Jan et al., 2017). In this case, the appendix contained all questions, but it was unclear how the questions related to the variables in the analysis. Another reporting issue is the describing, but never actually noting, the used questions (Fagundes et al., 2020; Buran Köse & Doğan, 2019; O'dea & Campbell, 2011; Servidio et al., 2018). An example of this could be a description like "include questions such as (…) and other relevant information" (Servidio et al., 2018). It is unclear what “other relevant information” means, so it is unclear what the researchers measured. The last reporting issue is the changing existing scales considerably without explaining in detail how

(Jarrar et al., 2022; Moningka & Eminiar, 2019); for instance, "it was adapted for the context of social media" (Moningka & Eminiar, 2019). More transparent reporting of changes to scales would, for instance, be “For use in the current study, “social media” replaced “Facebook” in the six items (…)” (Woods & Scott, 2016).

Another possible issue with robustness relates to the measurements; most of the literature relies on self-reports (84,51%). In these cases, researchers ask respondents to indicate their time or intensity of SNS use and ask them to rate their self-esteem level (Marengo et al., 2022); it is essential to be aware of these measurements' (possible lack of) reliability. For one, it could be doubted if self-report measures of social media use are representative of social media exposure. Experimental evidence would be more suitable for causal relationships, fortunately experiments were also available within the dataset (9,86%). The same could be wondered for self-esteem: is the measured self-esteem a reliable measure of actual self-esteem? Trying to solve this issue, one study attempted to test this relationship using embodied cognition (Shin et al., 2017); instead of asking respondents to rate their self-esteem level (which would be flawed), it was measured using the size of handwriting.

The second marker of robustness, transparency in data analysis, proved to be a complex concept to analyze. The reason for this is that the available papers only showed one final statistical model, meaning that it is unclear if and to what extent the analyses were altered after considering the data (a practice referred to as HARKing). What was possible to evaluate, however, was the proper reporting of statistics. As with the conceptualizations, the expectation should be that studies report all required statistics fully and correctly. However, in the sampled studies, there were various articles that did not do so. Examples of this were failures to report all path coefficients in mediation models, such as the total effect (8,45%) (Baturay & Toker, 2016; Bergagna & Tartaglia, 2018; Choi & Noh, 2019; Ozimek & Bierhoff, 2019; Romero-Rodriguez et al., 2020; Stapleton et al., 2017) and about 40% did not report necessary inferential information such as degrees of freedom (e.g., Acar et al., 2020; Cudo et al., 2019; Valkenburg et al., 2017). Besides the actual statistical results, preparatory steps before data analysis are missing from many papers; only a handful of papers report on how they dealt with missing scores (5,63%) (Fox et al., 2021; Schmuck et al., 2019; Tibber et al., 2020; Wang et al., 2017) or even fewer reported the results of a power analysis (5,63%) (Baturay & Toker, 2016; Busalim et al., 2019; Fox et al., 2021; Stapleton et al., 2017).

Another pattern within transparency that is interesting to note (though not necessarily a flaw) is the usage of control variables. Using control variables as confounding variables could be a researcher DF, in the sense that they can be used to create a significant effect when no significant effect is found in the main analysis, so it is essential to be critical when they are used (Simmons et al., 2011; Wicherts et al., 2016). Almost all papers use control variables like age or gender. However, a handful of studies (7,04%) base interpretations and conclusions on analyses with these variables, without priorly hypothesizing why these variables (and not others) are expected to serve as theoretical confounders or even discuss them in their theoretical framework (Demircioglu & Goncu Köse, 2020; Buran Köse & Doğan, 2019; Lui et al., 2017; Pettijohn et al., 2012; Yao et al., 2014).

P- & Z-curve

Figures 2a, 2b and 2c show the results of the p-curve analysis. These figures show three lines; the observed p-curve, the “null of no effect” and the “null of 33% power”. The observed p-curve shows the distribution of the p-values, the null of no effect shows the distribution of p-values if there was no effect. Lastly, the ‘null of 33% power’ shows the distribution of p-values if there was a population-level effect (with a sample powered at 33%). Comparing these lines can give an indication of evidential value; the observed p-curve should not be flatter than the “null of 33% power” line, which is not the case in any of the figures (The P-Curve App, 2017).

The p-curve analysis revealed that research on the effect of SNS use on self-esteem, so negative effects on self-esteem, had evidential value (k= 29, p= 0.0001; see Table 1 & Fig. 2a). The continuous tests show that evidential value is present for a full p-curve (Z=-15.76, p<0.0001) and a half p-curve (Z=-15.5353, p<0.0001). So, both indications for evidential value as formulated by Simonsohn, Nelson and Simmons (2013) are met. The estimate of the statistical power was 99% (90% CI [98%, 99%]).

For research theorizing the inverse relationship - the effect of self-esteem on SNS use - the analysis also indicated evidential value (k= 25, p<.0001; see Fig. 2b & Table 2). In this case, the continuous tests also show evidential value to be present for a full p-curve (Z=-16.88, p<0.0001) and a half p-curve (Z=-16.81, p<0.0001). So, also for this group, both indications for evidential value are met. The statistical power estimate was also 99% (95% CI [90%, 99%]). d

The last group, positive effects on self-esteem, does not show inadequate evidential value (k= 14, p = 0.61; see Fig.2c & Table 3). Also, the continuous tests show that evidential value is present for a full p-curve (Z=-4.14, p<0.0001) and a half p-curve (Z=-4.53, p<0.0001). The both tests indicate that evidential value is present. The estimated power for this group was 60% (90% CI [31%, 82%]).

Figures 3a, 3b, and 3c show the results of the z-curve analyses. The z-curve analysis showed for studies using negative effects on self-esteem (Figure 3a) showed that the expected replicability rate was 0.718 95% CI [0.492, 0.926]. The false discovery risk for this group was 24 95% CI [1, 98]. For self-esteem as an independent variable, the expected replicability rate was 0.718 95% CI [0.719, 1.000], and the false discovery risk was 4 95% CI [0, 30] (see Fig. 3b). The last group, positive effect on self-esteem, had an expected replicability rate of 0.573 95% CI [0.259, 0,881], and the false discovery risk was 34 95% CI [1, 100] (see Fig. 3c).

These statistics have considerably wide confidence intervals, which make drawing proper conclusions difficult.

Discussion

This paper aimed to answer the question of how strong the evidential value in the literature surrounding the relationship between social media use and self-esteem is and is there any indication for publication bias. This was done with a systematic literature review and both a p- and z-curve analysis with two sub-questions. The first sub-question was, "*To what extent is the literature surrounding the relationship between social media use and self-esteem conceptually transparent, based on the concepts of robustness and verifiability?*" Based on the systematic literature review, it can be concluded that much of the literature can benefit from more conceptual transparency in regard to reporting. The concepts of robustness and verifiability highlight common patterns of untransparent reporting; for instance, about 95% of the studies did not conduct a power analysis or did not report how they dealt with missing scores. The data also highlighted possible researcher degrees of freedom, like interpreting differences based on age and gender without priorly hypothesizing or not noting important statistical information. The lack of transparency was also evident in reporting measurements and instruments, with unclear descriptions of scales that make replication impossible. The lack of transparency was also evident in reporting measurements and instruments, with unclear descriptions of scales that make replication difficult.

The second sub-question, "*To what extent is p-hacking and publication bias present in the literature surrounding the relationship between social media use and self-esteem, based on a p- and z-curve analysis?*" was tested with the p- and z-curve analysis. The z-curve indicated sufficient replicability rates and false discovery risks; the confidence intervals were however too wide for conclusions. The p-curve gave evidence for a proper evidential value and sufficient estimated power for all groups. Based on these analyses, it can be concluded that all of the different groups of research, depending on the role of self-esteem, contained evidence for the absence of heavy p-hacking or publication bias.

To answer the main research question, this paper gives mixed results. Based on the p- and z-curve analyses, it could be concluded that the different groups contained evidential value. However, as will be discussed later, the p-curve also has its limitations, so is not hard evidence. The systematic literature review also highlighted conceptual flaws and some problematic patterns, which indicated methodological weakness, but it could also possibly decrease the evidential value.

These results mirror the complexity seen in the criticism of the current way of statistical testing (Amhrein et al., 2019; Ioannidis, 2005). Even though the analyses provide positive results, with seemingly sufficient evidential value, there are still some reasons for concern. A prime example concerns the large sample sizes surrounding these self-report studies, the null hypothesis will get rejected rather quickly; a slight variation in scores between survey items will lead to convincing evidence even with small effect sizes (Orben & Przybylski, 2019). In the context of this paper, about 85% of the literature used self-report studies. Since the p- and z-analyses used these statistics, it is important to keep this limitation in mind. It also begs the question if null-hypothesis testing is genuinely informative, since it is so influenceable by p-hacking and the choice of method. Alternatives like the Bayesian approach or putting a focus on effect size rather than p-values might be a solution (Masson, 2011).

But what about all of these contradicting findings concerning the role of self-esteem? There seem to be three types of findings: SNS use influences self-esteem, either negatively or positively, and self-esteem is a predictor of SNS use, which all contained sufficient evidential value. Earlier it was discussed that these contradicting findings could either result from the complexity of media effects or methodological and conceptual flaws (Orben & Przybylski, 2019; Valkenburg & Peter, 2013). The findings of this paper also deliver evidence for both theories. For one, the findings for all roles of self-esteem hold evidential value and reasonable replicability rates, so this might indicate the complexity of media effects. It seems that the relationship between self-esteem and social media use is not either positive, negative, or even a reverse effect, but depends on many other factors, since there is evidential value in all three effects. On the other hand, the systematic literature review did highlight some conceptual and methodological shortcomings concerning verifiability and robustness. This theory would be more problematic, as it decreases the evidential value of social media research. It is essential for the future of social media research to focus more on these underlying theories instead of testing (the same) specific media effects.

These results seem promising as the evidential value is proper, but this paper also has its limitations. For one, the literature used in this paper might not be exhaustive enough. One person did all the data collection and most of the coding. A second coder coded a small portion of the data; however, an ambitious project like a meta-analysis would benefit from more additional coders. Not all available literature was screened, but stopping rules like 'irrelevant articles' or 'only double articles' were used. Nevertheless, this does mean that the literature in this database is not fully representative of all available literature.

The replication aspects used in this paper, verifiability and robustness, were based on an extensive paper by Freese and Peterson (2017). However, so many more aspects could be covered for conceptual analysis. For one, the other aspects noted by Freese and Peterson, repeatability and generalizability could give concrete evidence for replicability as they actually intend to replicate findings. Secondly, there could be a focus on p-hacking techniques by using the data provided by researchers. This could create awareness and insight on statistical freedom and the dependency of significant results based on statistical choices. It was also found that there are different levels of transparency within these concepts, in the data availability, for instance. Some articles only make data available upon request, which begs the question if these are genuinely transparent. It does create awareness of open science as the concept of publishing data is noted, but immediately available data would be better, as anyone can contact a researcher requesting data.

Another limitation of this study derives from the incomplete reporting of studies in the database. The literature review discussed this incomplete reporting, but it also affects the p-curve analysis. Because many articles do not report t-values, degrees of freedom, or exact p-values, these had to be calculated by hand. These calculations might be based on incomplete and thus possibly incorrect data containing some errors. However, since the calculations were done based on the information given by the papers, they should theoretically be correct.

The p-curve analysis might be a proper tool to check the literature for evidential value; however, it does have its limitations. It has been demonstrated that the p-curve can incorrectly calculate effect size under certain conditions, such as when heterogeneity is moderate to large or when p-values are close to 0.05 (van Aert et al., 2016). Also, the p-curve analysis ignores non-significant results, so the used statistics might not represent all available literature. It makes sense with publication bias, but this does create an incomplete view of the relevant literature. Not only is this a limitation of the p-curve itself, but the z-curve would have been more complete if non-significant results were available. The comparison of the expected and observed discovery rate is a suitable indication of p-hacking and publication bias; however, this was not insightful in this paper since only significant results were used. It might be better to keep this in mind and code non-significant results when using these techniques for future research.

Replication research is still relatively new within communication science and is especially relevant. Since other scientific fields show that the evidential value can be lower than expected, it is fundamental to check our evidence as well. This paper might have positive results, but that does not immediately mean that all social media research contains the same evidential value. Therefore, future research must study more literature using similar techniques as this paper. This current paper has promising results for social media research but also found some weaknesses that are necessary to keep in mind. So, researchers should stay critical when reading papers, and don’t forget to publish the data.

References

Acar, I. H., Avcılar, G., Yazıcı, G., & Bostancı, S. (2020). The roles of adolescents' emotional problems and social media addiction on their self-esteem. *Current Psychology*. https://doi.org/10.1007/s12144-020-01174-5

Alfasi, Y. (2019). The grass is always greener on my Friends' profiles: The effect of Facebook social comparison on state self-esteem and depression. *Personality and Individual Differences*, *147*, 111–117. https://doi.org/10.1016/j.paid.2019.04.032

Anderson, C. A., & Warburton, W. A. (2012). The impact of violent video games: An overview. In W. Warburton & D. Braunstein (Reds.), *Growing Up Fast and Furious: Reviewing the Impacts of Violent and Sexualised Media on Children* (pp. 56–84). Australia: The Federation Press.

Andreassen, C. S., Pallesen, S., & Griffiths, M. D. (2017). The relationship between addictive use of social media, narcissism, and self-esteem: Findings from a large national survey. *Addictive Behaviors*, *64*, 287–293. https://doi.org/10.1016/j.addbeh.2016.03.006

Bartoš, F. & Schimmack, U. (2020, January 10). Z-Curve.2.0: Estimating Replication Rates and Discovery Rates. https://doi.org/10.31234/osf.io/urgtn

Baturay, M. H., & Toker, S. (2016). Self-Esteem Shapes the Impact of GPA and General Health on Facebook Addiction. *Social Science Computer Review*, *35*(5), 555–575. https://doi.org/10.1177/0894439316656606

Bergagna, E., & Tartaglia, S. (2018). Self-esteem, social comparison, and Facebook use. *Europe's Journal of Psychology*, *14*(4), 831–845. https://doi.org/10.5964/ejop.v14i4.1592

Berry, N., Emsley, R., Lobban, F., & Bucci, S. (2018). Social media and its relationship with mood, self-esteem and paranoia in psychosis. *Acta Psychiatrica Scandinavica*, *138*(6), 558–570. https://doi.org/10.1111/acps.12953

Boase, J., & Ling, R. (2013). Measuring Mobile Phone Use: Self-Report Versus Log Data. *Journal of Computer-Mediated Communication*, *18*(4), 508–519. https://doi.org/10.1111/jcc4.12021

Brandenberg, G., Ozimek, P., Bierhoff, H. W., & Janker, C. (2018). The relation between use intensity of private and professional SNS, social comparison, self-esteem, and depressive tendencies in the light of self-regulation. *Behaviour & Information Technology*, *38*(6), 578–591. https://doi.org/10.1080/0144929x.2018.1545049

Brown, J. (2018, January 5). *Is social media bad for you? The evidence and the unknowns*. BBC Future. Retrieved March 20, 2022, from https://www.bbc.com/future/article/20180104-is-social-media-bad-for-you-the-evidence-and-the-unknowns

Buran Köse, Z., & Doğan, A. (2019). The Relationship between Social Media Addiction and Self-Esteem among Turkish University Students. *Addicta: The Turkish Journal on Addictions*, *6*(1), 175–190. https://doi.org/10.15805/addicta.2019.6.1.0036

Burrow, A. L., & Rainone, N. (2017). How many likes did I get?: Purpose moderates links between positive social media feedback and self-esteem. *Journal of Experimental Social Psychology*, *69*, 232–236. https://doi.org/10.1016/j.jesp.2016.09.005

Busalim, A. H., Masrom, M., & Binti Wan Zakaria, W. N. (2019). The impact of Facebook Addiction and self-esteem on students' academic performance: A multi-group analysis. *Computers & Education*, *142*, 103651. https://doi.org/10.1016/j.compedu.2019.103651

Choi, D. H., & Noh, G. Y. (2019). Associations Between Social Media Use and Suicidal Ideation in South Korea: Mediating Roles of Social Capital and Self-esteem. *Health Communication*, *35*(14), 1754–1761. https://doi.org/10.1080/10410236.2019.1663466

Cingel, D. P., Carter, M. C., & Krause, H. V. (2022). Social media and self-esteem. *Current Opinion in Psychology*, *45*, 101304. https://doi.org/10.1016/j.copsyc.2022.101304

Cudo, A., Szewczyk, M., Błachnio, A., Przepiórka, A., & Jarząbek-Cudo, A. (2019). The Role of Depression and Self-Esteem in Facebook Intrusion and Gaming Disorder among Young Adult Gamers. *Psychiatric Quarterly*, *91*(1), 65–76. https://doi.org/10.1007/s11126-019-09685-6

Dean, B. (2022, January 5). *Instagram Demographic Statistics: How Many People Use Instagram in 2022?* Backlinko. Retrieved March 20, 2022, from https://backlinko.com/instagram-users

Demircioglu, Z. I., & Goncu Köse, A. (2020). Mediating effects of self-esteem in the relationship between attachment styles and social media addiction among university students. *Dusunen Adam: The Journal of Psychiatry and Neurological Sciences*, *33*, 8–18. https://doi.org/10.14744/dajpns.2019.00056

Dickersin, K. (1990). The Existence of Publication Bias and Risk Factors for Its Occurrence. *JAMA: The Journal of the American Medical Association*, *263*(10), 1385–1389. https://doi.org/10.1001/jama.1990.03440100097014

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., Lukito, J., Bier, L. M., Zhang, R., Johnson, B. K., Huskey, R., Schneider, F. M., Breuer, J., Parry, D. A., Vermeulen, I., Fisher, J. T., Banks, J., Weber, R., Ellis, D. A., . . . de Vreese, C. (2020). An Agenda for Open Science in Communication. *Journal of Communication*, *71*(1), 1–26. https://doi.org/10.1093/joc/jqz052

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication*, *12*(4), 1143–1168. https://doi.org/10.1111/j.1083-6101.2007.00367.x

Fagundes, L. S., Marot, T. A., & Natividade, J. C. (2020). Use of Instagram, Social Comparison, and Personality as Predictors of Self-Esteem. *Psico-USF*, *25*(4), 711–724. https://doi.org/10.1590/1413/82712020250410

Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association*, *105*(2), 203–206. https://doi.org/10.5195/jmla.2017.88

Fox, J., Vendemia, M. A., Smith, M. A., & Brehm, N. R. (2021). Effects of taking selfies on women's self-objectification, mood, self-esteem, and social aggression toward female peers. *Body Image*, *36*, 193–200. https://doi.org/10.1016/j.bodyim.2020.11.011

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*(6), 975–991. https://doi.org/10.3758/s13423-012-0322-y

Freese, J., & Peterson, D. (2017). Replication in Social Science. *Annual Review of Sociology*, *43*(1), 147–165. https://doi.org/10.1146/annurev-soc-060116-053450

Hawi, N., & Samaha, M. (2018). Identifying commonalities and differences in personality characteristics of Internet and social media addiction profiles: traits, self-esteem, and self-construal. *Behaviour & Information Technology*, *38*(2), 110–119. https://doi.org/10.1080/0144929x.2018.1515984

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Headspace. (n.d.). *How to deal with the negative effects of social media*. Retrieved March 20, 2022, from https://www.headspace.com/mindfulness/negative-effects-of-social-media

Holloway, F., Abdel-Khalek, A. M., Sagi, A., Muros, J. J., & Cofre-Bolados, C. (2016). *Self-Esteem: Perspectives, Influences and Improvement Strategies (Psychology of Emotions, Motivations and Actions)* (UK ed.). Nova Science Pub Inc.

Huang, C. (2021). Correlations of online social network size with well-being and distress: A meta-analysis. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *15*(2). https://doi.org/10.5817/cp2021-2-3

Instagram. (n.d.). *About Instagram*. Retrieved March 20, 2022, from https://about.instagram.com/

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jan, M., Soomro, S. A., & Ahmad, N. (2017). Impact of Social Media on Self-Esteem. *European Scientific Journal, ESJ*, *13*(23), 329. https://doi.org/10.19044/esj.2017.v13n23p329

Jarrar, Y., Awobamise, A., & Nweke, G. E. (2022). Social Communication Apprehension, Self-Esteem and Facebook Addiction Among University Students in Uganda. *Contemporary Educational Technology*, *14*(2), ep354. https://doi.org/10.30935/cedtech/11542

Koçak, O., İLme, E., & Younis, M. Z. (2021). Mediating Role of Satisfaction with Life in the Effect of Self-Esteem and Education on Social Media Addiction in Turkey. *Sustainability*, *13*(16), 9097. https://doi.org/10.3390/su13169097

Krause, H. V., Baum, K., Baumann, A., & Krasnova, H. (2019). Unifying the detrimental and beneficial effects of social network site use on self-esteem: a systematic literature review. *Media Psychology*, *24*(1), 10–47. https://doi.org/10.1080/15213269.2019.1656646

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C. M., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *Annals of Internal Medicine*, *151*(4), W. https://doi.org/10.7326/0003-4819-151-4-200908180-00136

Lim, A. J., Lau, C., & Li, N. P. (2021). The Moderating Role of Social Network Size on Social Media Use and Self-Esteem: An Evolutionary Mismatch Perspective. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.734206

Liu, Q. Q., Zhou, Z. K., Yang, X. J., Niu, G. F., Tian, Y., & Fan, C. Y. (2017). Upward social comparison on social network sites and depressive symptoms: A moderated mediation model of self-esteem and optimism. *Personality and Individual Differences*, *113*, 223–228. https://doi.org/10.1016/j.paid.2017.03.037

Malik, S., & Khan, M. (2015). Impact of facebook addiction on narcissistic behavior and self-esteem among students. *The Journal of the Pakistan Medical Association*, *65*(3), 260–263.

Marengo, D., Montag, C., Sindermann, C., Elhai, J. D., & Settanni, M. (2021). Examining the links between active Facebook use, received likes, self-esteem and happiness: A study using objective social media data. *Telematics and Informatics*, *58*, 101523. https://doi.org/10.1016/j.tele.2020.101523

Mares, M. L., & Woodard, E. (2005). Positive Effects of Television on Children's Social Interactions: A Meta-Analysis. *Media Psychology*, *7*(3), 301–322. https://doi.org/10.1207/s1532785xmep0703_4

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*(3), 679–690. https://doi.org/10.3758/s13428-010-0049-5

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

Moningka, C., & Eminiar, R. P. (2019). The Effect of Self-Comparison in Social Media on Self Esteem. *Advances in Social Science, Education and Humanities Research*, *494*, 383–389.

O'dea, B., & Campbell, A. (2011). Online Social Networking Amongst Teens: Friend or Foe? In B. K. Wiederhold, S. Bouchard, & G. Riva (Eds.), *Annual Review of Cybertherapy and Telemedicine 2011: Advanced Technologies in Behavioral, Social and Neurosciences - Volume 167 Studies in Health Technology and Informatics* (1st ed., Vol. 9, pp. 108–112). IOS Press.

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*(2), 173–182. https://doi.org/10.1038/s41562-018-0506-1

Ozimek, P., & Bierhoff, H. W. (2019). All my online-friends are better than me – three studies about ability-based comparative social media use, self-esteem, and depressive tendencies. *Behaviour & Information Technology*, *39*(10), 1110–1123. https://doi.org/10.1080/0144929x.2019.1642385

Pettijohn II, T. F., LaPiene, K. E., Pettijohn, T. F., & Horting, A. L. (2012). Relationships between Facebook Intensity, Friendship Contingent Self-Esteem, and Personality in U.S. College Students. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *6*(1). https://doi.org/10.5817/cp2012-1-2

Romero-Rodríguez, J. M., Aznar-Díaz, I., Marín-Marín, J. A., Soler-Costa, R., & Rodríguez-Jiménez, C. (2020). Impact of Problematic Smartphone Use and Instagram Use Intensity on Self-Esteem with University Students from Physical Education. *International Journal of Environmental Research and Public Health*, *17*(12), 4336. https://doi.org/10.3390/ijerph17124336

Rosenberg, M. (1965). Rosenberg Self-Esteem Scale. *PsycTESTS Dataset*. https://doi.org/10.1037/t01038-000

Saiphoo, A. N., Dahoah Halevi, L., & Vahedi, Z. (2020). Social networking site use and self-esteem: A meta-analytic review. *Personality and Individual Differences*, *153*, 109639. https://doi.org/10.1016/j.paid.2019.109639

Savage, J., & Yancey, C. (2008). The Effects of Media Violence Exposure On Criminal Aggression. *Criminal Justice and Behavior*, *35*(6), 772–791. https://doi.org/10.1177/0093854808316487

Schimmack, U. (2021, August 7). *Z-Curve: An even better p-curve*. Replicability-Index. Retrieved March 20, 2022, from https://replicationindex.com/2021/04/25/z-curve-an-even-better-p-curve/

Schivinski, B., Brzozowska-Woś, M., Stansbury, E., Satel, J., Montag, C., & Pontes, H. M. (2020). Exploring the Role of Social Media Use Motives, Psychological Well-Being, Self-Esteem, and Affect in Problematic Social Media Use. *Frontiers in Psychology*, *11*, 1–10. https://doi.org/10.3389/fpsyg.2020.617140

Schmuck, D., Karsay, K., Matthes, J., & Stevic, A. (2019). “Looking Up and Feeling Down”. The influence of mobile social networking site use on upward social comparison, self-esteem, and well-being of adult smartphone users. *Telematics and Informatics*, *42*, 101240. https://doi.org/10.1016/j.tele.2019.101240

Servidio, R., Gentile, A., & Boca, S. (2018). The mediational role of coping strategies in the relationship between self-esteem and risk of Internet addiction. *Europe's Journal of Psychology, 14*(1), 176–187. https://doi.org/10.5964/ejop.v14i1.1449

Shaw, L. H., & Gant, L. M. (2002). In Defense of the Internet: The Relationship between Internet Communication and Depression, Loneliness, Self-Esteem, and Perceived Social Support. *CyberPsychology & Behavior*, *5*(2), 157–171. https://doi.org/10.1089/109493102753770552

Shin, Y., Kim, M., Im, C., & Chong, S. C. (2017). Selfie and self: The effect of selfies on self-esteem and social sensitivity. *Personality and Individual Differences, 111*, 139–145. https://doi.org/10.1016/j.paid.2017.02.004

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

Stapleton, P., Luiz, G., & Chatwin, H. (2017). Generation Validation: The Role of Social Comparison in Use of Instagram Among Emerging Adults. *Cyberpsychology, Behavior, and Social Networking*, *20*(3), 142–149. https://doi.org/10.1089/cyber.2016.0444

Steinsbekk, S., Wichstrøm, L., Stenseng, F., Nesi, J., Hygen, B. W., & Skalická, V. (2021). The impact of social media use on appearance self-esteem from childhood to adolescence – A 3-wave community study. *Computers in Human Behavior*, *114*, 106528. https://doi.org/10.1016/j.chb.2020.106528

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. *The American Statistician*, *49*(1), 108–112. https://doi.org/10.2307/2684823

The p-curve app 4.06. (2017, November 30). The P-Curve App. Retrieved June 3, 2022, from http://www.p-curve.com/app4/

Tibber, M. S., Zhao, J., & Butler, S. (2020). The association between self-esteem and dimensions and classes of cross-platform social media use in a sample of emerging adults – Evidence from regression and latent class analyses. *Computers in Human Behavior*, *109*, 106371. https://doi.org/10.1016/j.chb.2020.106371

Trifiro, B. M., & Prena, K. (2021). Active instagram use and its association with self-esteem and well-being. *Technology, Mind, and Behavior*, *2*(3), 1–5. https://doi.org/10.1037/tmb0000043

Valkenburg, P. M., & Peter, J. (2013). The Differential Susceptibility to Media Effects Model. *Journal of Communication*, *63*(2), 221–243. https://doi.org/10.1111/jcom.12024

Valkenburg, P. M., Koutamanis, M., & Vossen, H. G. (2017). The concurrent and longitudinal relationships between adolescents' use of social network sites and their social self-esteem. *Computers in Human Behavior*, *76*, 35–41. https://doi.org/10.1016/j.chb.2017.07.008

Valkenburg, P. M., Pouwels, J. L., Beyens, I., van Driel, I. I., & Keijsers, L. (2021). Adolescents' social media experiences and their self-esteem: A person-specific susceptibility perspective. *Technology, Mind, and Behavior*, *2*(2). https://doi.org/10.1037/tmb0000037

Vall-Roqué, H., Andrés, A., & Saldaña, C. (2021). The impact of COVID-19 lockdown on social network sites use, body image disturbances and self-esteem among adolescent and young women. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *110*, 110293. https://doi.org/10.1016/j.pnpbp.2021.110293

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting Meta-Analyses Based on p-Values. *Perspectives on Psychological Science*, *11*(5), 713–729. https://doi.org/10.1177/1745691616650874

Walton, A. G. (2021, December 10). *6 Ways Social Media Affects Our Mental Health*. Forbes. Retrieved March 20, 2022, from https://www.forbes.com/sites/alicegwalton/2017/06/30/a-run-down-of-social-medias-effects-on-our-mental-health/?sh=4b71783b2e5a

Wang, P., Wang, X., Wu, Y., Xie, X., Wang, X., Zhao, F., Ouyang, M., & Lei, L. (2018). Social networking sites addiction and adolescent depression: A moderated mediation model of rumination and self-esteem. *Personality and Individual Differences*, *127*, 162–167. https://doi.org/10.1016/j.paid.2018.02.008

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*, 1–12. https://doi.org/10.3389/fpsyg.2016.01832

Wilcox, K., & Stephen, A. T. (2013). Are Close Friends the Enemy? Online Social Networks, Self-Esteem, and Self-Control. *Journal of Consumer Research*, *40*(1), 90–103. https://doi.org/10.1086/668794

Williams, N., PhD. (2021, December 7). *How is Social Media Affecting Mental Health?* News-Medical.Net. Retrieved March 20, 2022, from https://www.news-medical.net/health/How-is-Social-Media-Affecting-Mental-Health.aspx

Woods, H. C., & Scott, H. (2016). #Sleepyteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem. *Journal of Adolescence*, *51*(1), 41–49. https://doi.org/10.1016/j.adolescence.2016.05.008

Yao, M. Z., He, J., Ko, D. M., & Pang, K. (2014). The Influence of Personality, Parental Behaviors, and Self-Esteem on Internet Addiction: A Study of Chinese College Students. *Cyberpsychology, Behavior, and Social Networking*, *17*(2), 104–110. https://doi.org/10.1089/cyber.2012.0710

Appendix

Table 1

P-curve statistics for self-esteem as dependent variable group

	Binomial tests	**Continuous Tests Full curve (p's <0.05)**	**Continuous Tests Half curve (p's <0.35)**
Studies contain evidential value	p = 0.0001	Z=-15.76, p <0.0001	Z=-15.53, p <0.0001
Studies' evidential value, if any, is inadequate	p = 0.9835	Z=10.65, p >0.9999	Z=13.98, p >0.9999

Table 2

P-curve statistics for self-esteem as independent variable group

	Binomial tests	**Continuous Tests Full curve (p's <0.05)**	**Continuous Tests Half curve (p's <0.35)**
Studies contain evidential value	p = 0.0001	Z=-16.88, p <0.0001	Z=-16.81, p <0.0001
Studies' evidential value, if any, is inadequate	p = 0.9982	Z=12.23, p >0.9999	Z=43.7, p >0.9999

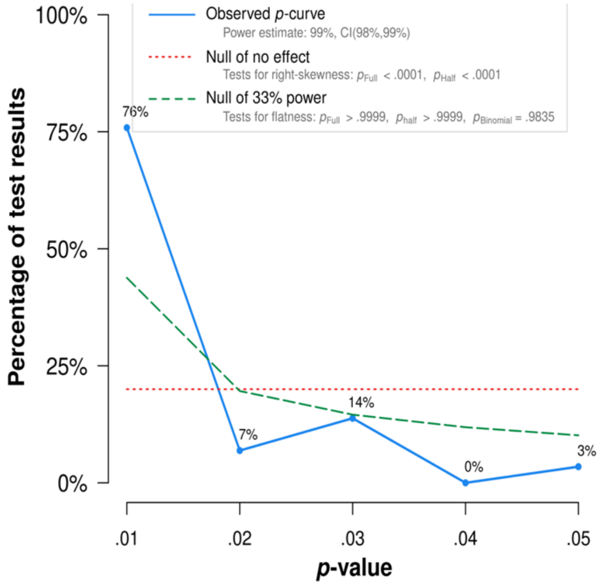
Table 3

P-curve statistics for positive effects on self-esteem group

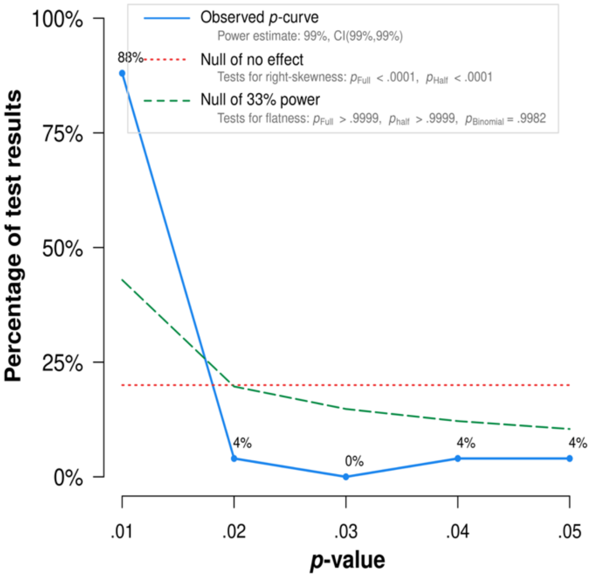
	Binomial tests	**Continuous Tests Full curve (p's <0.05)**	**Continuous Tests Half curve (p's <0.35)**
Studies contain evidential value	p = 0.0898	Z=-4.14, p <0.0001	Z=-4.53, p <0.0001
Studies' evidential value, if any, is inadequate	p = 0.6072	Z=1.49, p = 0.9313	Z=4.71, p >0.9999

Figure 2

a) Influence of SNS use on self-esteem



b) Influence of self-esteem on SNS use



c) Positive influence of SNS use on self-esteem

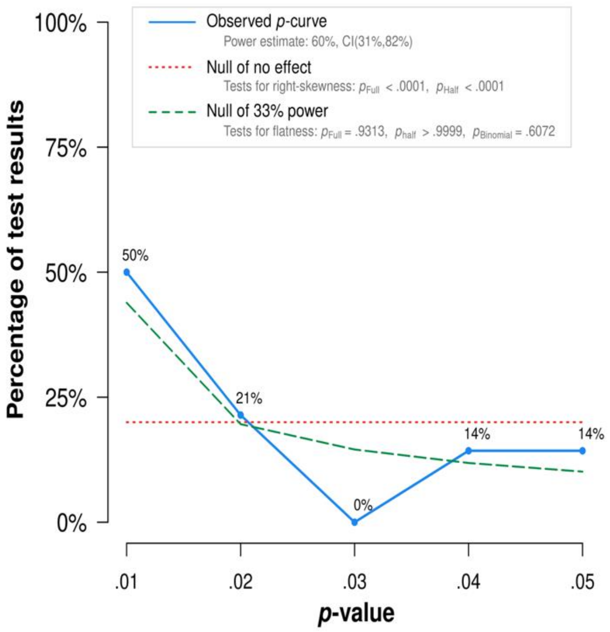
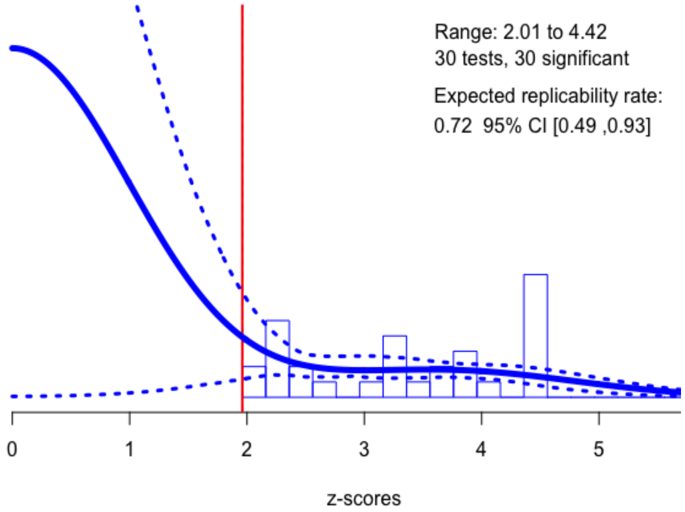
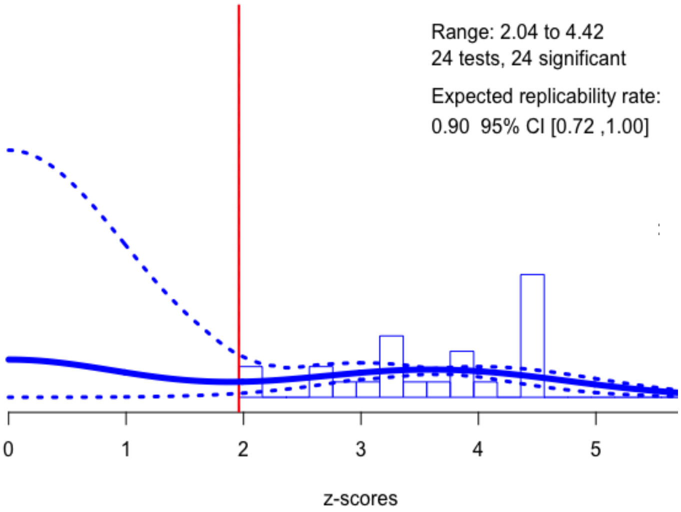


Figure 3

a) Influence of SNS use on self-esteem



b) Influence of self-esteem on SNS use



c) Positive influence of SNS use on self-esteem