



UPPSALA
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Social Sciences 209*

Challenges when Generalizing Psychological Measurements across Populations

*Applications in Machine Learning and Cross-Cultural
Comparisons*

MATHIAS BERGGREN



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2023

ISSN 1652-9030
ISBN 978-91-513-1720-5
URN urn:nbn:se:uu:diva-496532



Dissertation presented at Uppsala University to be publicly examined in Sal IX, Universitetshuset, Biskopsgatan 3, Uppsala, Wednesday, 5 April 2023 at 09:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Mikael Hjerm (Sociologiska institutionen, Umeå University).

Abstract

Berggren, M. 2023. Challenges when Generalizing Psychological Measurements across Populations. Applications in Machine Learning and Cross-Cultural Comparisons. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences* 209. 77 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-1720-5.

In order to ascertain the validity and applicability of psychological theories, models, and measurements, it is important to examine their generalizability across different assessment situations. In this thesis, I examine how the application of measures outside of their initial domain may cause complications. This is applied to two fields where such considerations of generalizations may be especially beneficial: machine learning models and cross-cultural comparisons. Paper I explored whether text-based machine learning models of personality with a broad set of predictors, or models based on a set of more constrained but more psychologically meaningful predictors, better predicted personality in one of two text domains. The former models provided equal or superior prediction in the same domain in which it was trained compared to the latter models, but equally poor or poorer prediction in the other domain. Paper II reexamined the results of an article that, like the cross-cultural studies re-examined in Paper III, found that over time and across states in the U.S., higher gender equality was associated with larger gender differentiation, here in names given to children. Re-analyses showed that there was no such systematic association across time, and that the differentiation across states was confounded with a more strongly associated cultural/language predictor. Paper III re-examined multiple studies that have assessed that association across countries. Here, it was shown that cultural differences, as indicated by cultural regions, other measures such as individualism, and data quality indicators, better explained the variation in differences across countries. When controlling for cultural/language regions, the association with gender equality disappeared or, sometimes, reversed. These results indicate the degree to which different cultural factors are interrelated, and suggests the need for complementary methods. In conclusion, this thesis exemplifies the importance of considering how models and measures may interact with and generalize across situations. This is true whether it supports greater generality or situational specificity of different psychological measures.

Keywords: personality, machine learning, culture, gender differences

Mathias Berggren, Department of Psychology, Box 1225, Uppsala University, SE-75142 Uppsala, Sweden.

© Mathias Berggren 2023

ISSN 1652-9030

ISBN 978-91-513-1720-5

URN urn:nbn:se:uu:diva-496532 (<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-496532>)

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

- I. Berggren, M., Kaati, L., Pelzer, B., Stiff, H., Lundmark, L., & Akrami, N. (manuscript submitted for publication). The generalizability of machine learning models of personality across two text domains.
- II. Berggren, M. (in press). No evidence of a gender-equality paradox in gendered names: Comment on Vishkin, Slepian, & Galinsky (2022). *Social Psychological and Personality Science*.
- III. Berggren, M., & Bergh, R. (manuscript). A re-assessment of some cross-cultural associations with psychological gender differences.

Reprints were made with permission from the respective publishers.

The contribution of Mathias Berggren to the studies was as follows:

Paper I: Co-planned and co-designed the study, analyzed the data, and wrote the manuscript with contributions from the co-authors.

Paper II: Sole author. Helpful discussions and comments were provided by my supervisors.

Paper III: Planned and designed the study, gathered and analyzed the data, and wrote the manuscript with contribution from the co-author.

Contents

Introduction.....	7
Psychological Constructs	8
Challenges with Assessing Causality.....	9
Reliability, Validity, and Challenges Capturing Error	10
Method Summary	14
Generalizability and Replicability.....	14
A Cultural Bias in Psychology?	17
Summary and Applications	19
Machine Learning Models of Personality	19
Machine Learning Basics	19
Model Misspecification	21
Model Complexity and Model Robustness.....	22
Limits and Opportunities of Machine Learning Models.....	24
Cross-Cultural Assessments of Psychological Gender Differences.....	25
Theoretical Background.....	26
Some Possible Cross-Cultural Confounds	30
Aims.....	32
Paper I.....	33
Background	33
Method	34
Results.....	34
Conclusions	38
Paper II.....	39
Background	39
Method	42
Results.....	42
Conclusions	44
Additional Considerations.....	45
Is the measure of gender differences well-validated?.....	45
How influential can bias in earlier years be?.....	47
Can one remove the names of children of foreign-born parents?.....	48
Dependence over time in the English and Welsh dataset?	49
Paper III	51

Background	51
Method	53
Results	55
Conclusions	58
General Discussion	59
Summary of Key Results.....	59
Models may Interact with the Context	61
Measures may be Cross-Culturally Confounded.....	62
Limitations	64
Future Directions.....	66
Conclusions	68
Acknowledgements.....	69
References.....	71

Introduction

Psychology often aims to describe universal patterns of human behavior, yet is perhaps just as often limited to particular populations or operationalizations in its empirical examinations (e.g., Henrich et al., 2010a; Yarkoni, 2022). Following the Replication Crisis, there has been an increased recognition of the need to examine the reliability of statistical results and the validity of their associated theories, beyond the initial inquiry (e.g., Shrout & Rodgers, 2018).

This thesis deals with one perhaps less examined aspect of psychological results: the degree to which they are generalizable. That is to say, when psychological theories or constructs are generalized outside of the domain of their initial conception, how do they hold up to this change in scenery. The entirety of this question is beyond the scope of this thesis; the common thread throughout, however, is to advocate for the importance of considering the complexity of the surrounding context when interpreting results.

More specifically, the papers contained in this thesis all start out with some model of human behavior informed by a particular context. Then it is examined how the application of those models to other contexts raises additional challenges. Paper I deals with machine learning methods for predicting personality. Studies in this field have sometimes made quite strong assertions regarding the applicability of- and theoretical contribution provided by such models. In this paper, we examine models' generalizability, and discuss ways in which a greater consideration of the context may help the field to fulfill these aspirations.

Papers II-III then deal with generalizability across cultural contexts when measures and models of human behavior developed in the West are applied to other cultures. In these papers, we examine how a bias towards a Western perspective may cause complications when comparisons are made between populations. More specifically, it deals with the measurement of gender differences, and the way this can be influenced by cultural differences.

Before presenting these papers, however, we require some background information regarding psychological constructs and their measurements, and accompanying challenges when trying to generalize results across contexts. We start by discussing how psychological constructs are conceptualized, and some challenges that exist with capturing the causality often construed in such constructs. Here, it is argued that these challenges require researchers to consider how predictions from their theories and accompanying conceptualization

of psychological constructs hold up under differing contexts, and that a limited consideration of such contexts may lead to theories to appear more general than what has been established. Next, we turn to some additional measurement challenges that may make it difficult to parse different conceptualizations of psychological constructs from each other, and that may cause results to become more situationally dependent.

With this introduction to the challenges with separating different psychological explanations for results from each other, we next turn to the importance of examining generalizability as a way of overcoming these challenges. Here, it is discussed how the varying of situations and assessment methods may give further proof of validity of psychological explanations for certain results, beyond how replicability establishes the reliability of those results.

We subsequently turn to one particular way in which psychological results may not have been sufficiently examined for generalizability: the degree to which they are culturally specific to (amongst other things) a Western context. An insufficient consideration of this Western cultural specificity may have particular problems when psychological theories are applied in other cultures, and when results on Western measures are compared across cultures.

Finally, we apply and develop these ideas when discussing two main areas considered in this thesis: machine learning models of personality, and the examination of psychological gender differences across different populations.

Psychological Constructs

Psychology is ripe with constructs meant to capture some predictable pattern in human thoughts, feelings, and behaviors: personality traits, general intelligence, conformity, motivation, reasoning, attachment, et cetera. Definitions of psychological constructs abound (see e.g., Slaney, 2017, p.3-4), but one classical definition of constructs is as “*some postulated attribute of people, assumed to be reflected in test performance*” (Cronbach & Meehl, 1955, p.283). To clarify the special status of hypothetical constructs, MacCorquodale and Meehl (1948) further define that they “*involve the supposition of entities or processes not among the observed*” (p.106-7). Thus, constructs typically do not just refer to the observable phenomena (e.g., “answers on items 1-24 correlate positively”), but function as a theoretical and explanatory model for those phenomena (e.g., “one broad underlying extraversion factor causes the commonality between items 1-24”).

This illustrates constructs’ close connection to psychological theories. Theories function to explain observed phenomena, and constructs are the hypothetical entities or processes referenced to within such theories to help with that explanation. A person is argued to act a certain way because they are “extraverted”, they get high scores on an IQ test because they are “intelligent”, they work hard because they are “motivated”, and so on. But this also means

that such constructs contain something beyond what is observed (and, conversely, they also focus on some aspects of what is observed, while leaving other things aside).

Thus, a psychological construct, like a psychological theory, cannot be fully validated from just one type of measurement (e.g., Campbell & Fiske, 1959). Constructs are concerned with describing the, in themselves, unobservable psychological characteristics of people. Thus, no measurement is a direct one-to-one reflection of the construct. In this separation between constructs and its indicators, other factors can come into play. Similarly, as the full extent of the construct is not captured by any one indicator, varying constructs may explain any finite number of assessed indicators. Thus, psychological constructs entail a continual examination of the predictions derived from the conceptions of those constructs, to provide further proof of their validity. Concurrently, different explanations for the same phenomena need to be entertained. In conducting such examinations, however, there are several measurement challenges which may limit researchers' ability to discern between different theories and construct conceptualizations. We turn to discuss some such challenges in the next section.

Challenges with Assessing Causality

Psychological theories tend to be causal in nature. Indeed, as previously discussed, psychological constructs are posited as an explanation of behavior, and then often as a causal explanation. Depending on the construct examined, however, measuring the effects of that construct in a way which allows for causal conclusions may be problematic. If the understanding of personality, for example, is as a biological and highly stable characteristic of people that causes behavior (see McCrae & Costa, 2008), then it is naturally difficult to do an experimental manipulation on it to examine whether peoples' behavior changes as predicted. Instead, one way in which the "primacy" of personality traits is examined is by establishing their reliability across time (e.g., Rantanen et al., 2007). If the underlying tendency towards different personality traits cannot change, then this provides some support that they are a cause, rather than an effect of behavioral differences between individuals. However, this does not as directly support the hypothesized causal nature of personality traits as would an experimental manipulation, and thus does not conclusively support the exogenous "biological entity" perspective compared to other ones. A competing perspective is of personality and abilities as emergent properties (see e.g. Baumert et al., 2017; Borsboom et al., 2003; Cramer et al., 2012; van der Maas et al., 2006, Schmittmann et al., 2013). Further, as some variation in the trait is typically allowed, it becomes difficult to delineate when a causal effect from the personality trait to individual behaviors is falsified in favor of, for example, a causal effect from individual behaviors to the personality trait.

This latter case is the direction predicted by emergent perspectives, where personality traits act as summaries of the logical interdependencies between related behaviors, thoughts, and feelings.

Consequently, many examinations of personality, and of psychological constructs more broadly, rely on correlational data. Such correlations from cross-sectional data, as is often repeated, can establish association, but not finally account for causation. This both because of the problem with establishing a causal direction, as discussed above, but also because the correlation between two variables need not reflect a causal relationship between those variables at all. Other unexamined variables may instead affect both of the examined variables, and thus drive the correlation. Had the construct been possible to manipulate (unequivocally), then the question of causality could be addressed (although, causality in the other direction could still exist as well). With correlations, it may be possible to derive at more or less plausible causal models, for example, in mediation analysis (e.g., MacKinnon et al., 2007), by examining the strength of association between different variables. However, this may require a thorough examination of different variables to exclude other effects, and, even then, if there is error in variables, this may confound results (see next section).

These points will, in some form or another, be relevant for all papers in this thesis. In Paper I, the issue is that when constructing machine learning models, the causal relationship between variables is typically poorly understood. Instead, the strongest and most reliably correlated variables are identified and used to construct the best-fitting model in the examined situation. However, even if variables (or combinations of variables) are strongly correlated in a situation, there is no guarantee that they are closely associated in a causal sense, or that they generalize reliably across situations. In Paper II, and especially Paper III, the “correlation does not equal causation” dictum, may be even more central. These papers reply to a causal hypothesis; that increased gender equality leads to increased psychological and behavioral gender differences (see Schmitt et al., 2008; see also e.g., Mac Giolla & Kajonius, 2019). However, this hypothesis rests, to a large part, on correlations across countries (an exception is the article responded to in Paper II). Such correlations appear especially vulnerable to confounding factors that may exist between cultures.

Reliability, Validity, and Challenges Capturing Error

The concepts of reliability and validity, used within psychology to delineate how different sources of error affect results, is similar to those of variance and bias within machine learning and statistics (e.g. James et al., 2013, Ch. 2.2.2; although validity can also be seen as a broader concept than bias). A common way to illustrate the importance of reliability/variance and validity/bias of measurements is with a metaphorical archery target. Ideally, to capture the

intended psychological construct, one should hit a bullseye with each shot. For this, however, both high reliability and validity is needed. With lowered reliability, which is typically understood as non-systematic, or random, measurement error, shots get increasingly varied around the target. With lowered validity, there is increasing bias in the shots, so that they no longer appear to target the bullseye (see e.g. McCrae et al., 2011 for a discussion of reliability and validity in relation to personality assessments). To exemplify with the personality trait of Extraversion, if no item of an Extraversion test measures it perfectly, but the error in items cancel out in the test as a whole, then reliability is not perfect, but there are no validity-problems. This is what is assumed when factor-analysis is applied to personality tests (see e.g., Russell, 2002). Should there be some systematic deviation, though, say if items meant to capture Extraversion also tend to capture the separate personality trait Agreeableness, then the test has validity problems.

Validity issues have previously been brushed upon in connection to confounding factors and the challenges of assessing causality. Here, we will therefore focus on some challenges with reliability, as well as challenges with measuring and controlling for random and systematic error. In self-assessment scales, reliability is typically assessed with Cronbach's alpha, which is a measure of scale items' consistency (Cronbach, 1951). Consistency can be expected to deteriorate for several sorts of data quality reasons, and is therefore useful to consider. For example, if some items do not work as intended, or if people consider items less carefully, then their consistency may be expected to decrease. However, Cronbach's alpha, and other measures of reliability, are not universal measures of data quality (see e.g. Schmidt et al., 2003; Sijtsma, 2009), and several other ways of controlling for or assessing data quality exists (see e.g., Curran, 2016).

The use of Cronbach's alpha stems from the classical test theory understanding of response error as error added independently to the variable of interest – that is, error which decreases reliability but not validity. When this holds true, repeated measurements can be averaged to provide an unbiased measurement of the underlying variable, with lower error than a singular assessment. This is for example reflected in how Cronbach's alpha increases with the number of items if the covariation between items otherwise remains the same (see e.g. Streiner, 2003). Over time, a more complex consideration of error has materialized, however. This is for example illustrated by the changing understanding of how careless responding can enter into survey data and affect results. An overview of these issues is presented by Curran (2016). Careless responding is more difficult to control for than pure independent random error, as the “method of carelessness” that is employed may lead to those careless responses manifesting in different ways – that is, they may cause different types of bias as well as random error. Thus, any single measure of problematic data can at best only identify a subset of such responses (e.g. Meade & Craig, 2012). For this reason, Curran (2016) identify several indices of data

quality that may be used in tandem to identify different subsets of careless responses.

When it comes to Cronbach's alpha, for example, this will be a less direct predictor of careless responses if those responders can remain consistent between scale items even for spurious reasons. Assume, for instance, that participants randomly pick a response option for the first item in a scale, and then recognize that later items measure similar things, and therefore provide similar answers. The easier it is to remain consistent for such spurious reasons, the more consistent will even be careless responses that do not measure the trait of interest, and the less Cronbach's alpha will be affected. The tendency to remain consistent may depend both on how easy it is to recognize that different items measure roughly similar things, as well as how motivated responders are to appear to answer thoughtfully. Thus, Cronbach's alpha and other measures, may only be indirect measures of how data quality decreases. In the Supplementary Information of Paper III, this is illustrated with a simulation where careless responses increase to high levels, although Cronbach's alpha does not drop much, as there remains careful responders who help it remain high. This model also illustrates one way in which poor responses can affect results: by attenuating the measured difference between groups (see e.g. Alwin & Krosnick, 1991; Curran, 2016; see also Loken & Gelman, 2017), which drops much more in these simulations. In Paper III, we use Cronbach's alpha as our main indicator of data quality, as this is the measure available for all self-assessment studies we re-examine. As it can be expected to covary with different types of data quality aspects, it is an interesting measure to examine. However, as already mentioned, it will not be capable of capturing all types of data quality, and, simultaneously, although it may covary with different types, more direct measures might be possible if one can identify the reason for some specific differences in data quality variation.

The issue that poor measurements and/or poor responses can cause the magnitude of measured group (and individual) differences to diminish as response quality deteriorates, also causes problems when estimates between different variables are compared. For instance, this would happen if two independent variables compete in predicting a dependent variable within a multiple regression model. The more strongly correlated variable then tend to win out over the other if they otherwise explain similar parts of the variation in the dependent variable. Commonly then, the winning variable is treated as the better predictor of the dependent variable. If a causal interpretation is added, it may also be seen as the most directly associated of the variables, and in some cases, as the true (or at least truer) cause of the dependent variable. Further, if both variables have independent contributions for the prediction of the dependent variable, the constructs they are meant to measure are also assumed to have independent contribution. However, for these conclusions to hold, the variables should also be measured with a similar level of precision (see e.g.,

Berggren, 2020). If they are not, then that imprecision may artificially attenuate the association to different degrees, possibly making it so that a weaker associated but more precisely measured variable gets favored over a more strongly associated but less precisely measured variable. Similar problems arise, for example, with restriction of range in variables (e.g., Sackett et al., 2007), and the commonly described independent additive error-in-variable issue (e.g., Carroll et al., 2006).

For the papers in this thesis, these factors are first of relevance for the machine learning methods in Paper I. Here, the best predictive model of a personality trait should depend on the particular population and situation under which the prediction is conducted. This is first because different situations may vary in the degree to which they are conducive for different behaviors (which function as the predictors). However, it will also be affected by variations in measurement-quality across populations and situations, as the situation may make different behaviors more or less difficult to measure. As an example, measuring indicators of peoples' tendency towards Conscientiousness, that is how motivated, industrious, and organized one is (e.g., McCrae & Costa, 2008), may be harder during a job-interview than during normal working hours. This may be because everyone tries to appear more Conscientious during a job-interview, leading to a restriction of range in the behaviors. This tendency may also cause the behaviors to depend on other factors than it would during normal working hours. Here, for example, such factors may include how socially adept a person is, or how desperate they are for the job. Additionally, because the assessment takes place during a shorter period of time, any measurement will be restricted to a coarser measurement than would be, for example, a measurement over normal working hours. All in all, then, there are many reasons why personality traits and other psychological constructs may be differentially predicted in different situations, for conceptual as well as measurement reasons. In Paper I, we examine how personality is predicted from text in two different situations, online messages, and personal essay. As discussed in relation to the results of that paper, the personal essays may have been more conducive for finding personality patterns. In addition, we consider how different types of models may vary across situations to a greater or lesser degree as a result of how variables, and the training method of the models, may be differentially affected by such situational factors. Specifically, we examine how differences in the degree of model complexity may yield predictions that are more or less specific to a certain situation, predicting that higher dimensional models may be particularly conducive for finding predictions that work well in one situation, but may not necessarily generalize across situations. Here, we examine whether a broad set of such predictors, or a (more) constrained set of more psychologically meaningful, and therefore perhaps more centrally and generally important, predictors yield stronger predictors both when the model is applied on the same situation in which it was trained, as well as when the situation changes.

Measurement challenges are also highly relevant for Paper III. Here, we reassess previously found variation in gender differences between different countries (first observed by Costa et al., 2001). If response/data quality differs between those countries, then this may thus be one important factor for those differences. These reassessments may indicate that data quality differs between countries in a way which may produce (some) differences. However, a complete accounting for those data quality differences between countries will be beyond the scope of this paper, due to the various ways in which such poor data quality may manifest itself. Future studies may assess data quality in greater detail.

Method Summary

Methodological issues may cloud interpretations of results and lead to incorrect conclusions. Thus, it is important to consider such methodological confounds when interpreting results, and to further corroborate one's findings with new designs that may avoid such issues, or at least controls for them. Controlling for methodological confounds may be difficult, however, as it may be difficult to identify error in responses and the most appropriate model. Such methodological factors may also lead to the same problem repeating itself in direct replications, meaning that further similar examinations will not come to terms with these issues. For this reason, in the next section, we discuss the importance of examining the generalizability of psychological theories and results.

Generalizability and Replicability

The Replication Crisis in psychological science (and more broadly throughout the sciences) is the recognition that many classical psychological results cannot be replicated reliably (see e.g., Open Science Collaboration, 2015; Loken & Gelman, 2017; Shrout & Rodgers, 2018). One explanation for this crisis is the use of questionable research practices. This involves researcher's degrees of freedom (Simmons et al., 2011), that is: flexibility in data collection and analysis which increases the likelihood of finding spuriously significant effects (see also Gelman & Loken, 2013 for discussion on how this can be a problem even when there are no conscious such efforts from the side of researchers). Examples include whether to collect or not collect additional observations after seeing results, and choices between which variables to analyze or to control or not control for. With the recognition of a Replication Crisis, there has been a push to improve the way psychological science is conducted,

with several suggested areas of improvement both for when the science is conducted, as well as for the systematic evaluation, peer-review, and establishment of the reliability of results through replication (Asendorpf et al., 2013).

Yet, it has also been noted that replication of results, although an important part of science, is not enough to ascertain the validity of researchers' hypotheses (e.g., Yarkoni, 2022; see also Asendorpf et al., 2013). Results may indeed be highly replicable, while providing very little information to support one theory over another. Worse, they may provide a highly biased or one-note picture of the relationship between variables specific to the particular examination method employed (see e.g., Yarkoni, 2022). Thus, the generalizability of psychological theories and models over different types of assessment methods and with different populations and situations also needs to be examined. This type of examination is sometimes called conceptual replication (e.g., Fabrigar & Wegener, 2016, p.68), in which case what we simply call replication is termed "direct replication".

The danger when the support for a certain theory relies heavily on a specific assessment method is that the predictions from that theory may not hold when other types of assessments are employed (for example, as discussed in relation to Paper III, results across countries may not replicate as predicted across time). This is always problematic, but perhaps especially so when the results were already found before the introduction of the hypothesis. The reason for this danger has to do with what we touched upon regarding psychological constructs and measurements: several different psychological (or non-psychological) explanations may explain the same type of results. Thus, when an explanation is given to results already established, this may be more indicative of the possibility of formulating an engaging explanation, rather than the predictive power of that explanation. This is the important distinction between exploratory and confirmatory hypothesis testing (see e.g., Nilsen et al., 2020; Nosek et al., 2018; Simmons et al., 2011) – exploratory testing being conducted without strong predictions from any one theory, while confirmatory testing is conducted to examine whether well-established predictions from one or more theories hold up to scrutiny.

Even if results are found through confirmatory testing, however, this may not be enough if multiple competing accounts can explain the same phenomenon. That is, if the method meant to test one's hypothesis is conducted so the predicted results could also come about due to other competing hypotheses. Then the test becomes unable to support one theory over the other. To provide more robust proof that one's results depend on one's theory, it has therefore been argued that one should conduct what has been called severe tests (see e.g., Mayo, 1991), or tests general enough that they correspond to the varying situations under which the theory is supposed to hold true (Yarkoni, 2022). That is to say: it is not enough to test any prediction from one's hypothesis against a competing null-hypothesis which does not predict an effect in that direction, and consider it supported if one gets results unusual enough under

the null-hypothesis (the standard null hypothesis significance testing procedure; see e.g., Gigerenzer, 2004). Instead, researchers also need to consider the results to expect under other competing hypotheses, and find tests specific (severe) enough that a result in the predicted direction under their hypothesis would not also happen under other hypotheses. This could perhaps be done by specific enough singular tests; however, it may be difficult to completely separate competing explanations in any one test. Thus, examining the generalizability of one's explanation, by continuously testing new predictions from one's theory, may be the more approachable way. The severe tests are then the sum of the tests from those varying predictions (we may consider this a more temporally drawn-out alternative to the random-effects models favored by Yarkoni, 2022). Generalizability thus becomes one way to further test the robustness of researcher's results, and provide more specific proof for certain theories over others.

From the above discussions, one way in which generalizability provides additional proof of concept over that of replicability is thus that it provides stronger proof that the reason for the results has to do with one's hypothesis, compared to other explanations. Thus, generalizability works by providing further support for the validity of one's theories for the observed results. This is beyond how replicability establishes the reliability of results. Such replication-reliability is necessary to establish that there even is some statistical pattern worthy of consideration. However, once this has been done, examining generalizability provides additional crucial proof of concept. Thus, replicability alone is not enough to establish the validity of one's theories. This point will be of importance in Paper II-III, which respond to a highly replicated result, which, however, has been replicated in methodologically limited ways which do not directly correspond to the causal predictions that it relies on.

In addition to providing proof for the validity of certain theories over others, generalizability also naturally examines how valid one's predictions are in different situations. That is, replicability, once again, may show that one's predictions hold in a certain situation, but generalizability examines how far those predictions hold. This is also of importance to the machine learning models of personality examined in Paper I. Examining the generalizability of such models should provide researchers with an improved understanding of the generality or specificity of different predictors of personality. In addition, it shows how influential different situational factors are for the variables that come out as important predictors for personality. Therefore, it may also provide information about how and how much different expressions of personality vary between situations.

A Cultural Bias in Psychology?

One particular way in which the generalizability of psychological results appears to have been underexamined is in considering their possible cultural specificity. To see how culture may influence results, consider the following statement similar to one Conscientiousness item in the NEO-PI-R Five Factor instrument (Costa & McCrae, 1992): “I do not care that much about civil duties such as voting”. This item will carry quite different meanings in a country with free democratic elections and high trust in its institutions, versus one without democratic elections, or with high corruption and mistrust in the system. The identification of voting as a civil duty that a high-Conscientiousness person is likely to agree with seems to presuppose the former, or at least to be more relevant an indicator of individual differences in Conscientiousness in such a country. Variation in how relevant items are for people in different countries, including in subtler ways than the example above, may thus complicate cross-cultural comparisons. Indeed, personality items appear to show lowered validity or different factor structures as one moves outside of the West (e.g., Laajaj et al., 2019; Saucier & Goldberg, 2001).

The general tendency in psychology to have a specific cultural bias in its methods and assessments has been called the WEIRD-bias (Henrich et al., 2010a). This draws attention to the fact that psychological assessments are commonly done with a Western, Educated, Industrialized, Rich, and Democratic point of reference (Henrich et al., 2010a; Henrich et al., 2010b). The WEIRD acronym is further meant to indicate that this is a very specific cultural context that may not be representative of other populations.

From the previous discussions, we see that there may be issues when measures developed within the Western context are used in cross-cultural research, as those measures may have been constructed in accordance with cultural conceptions from the West (i.e., have a WEIRD-bias) – as was the case with the voting question, for example. The issues that may arise when Western measures are applied cross-culturally can take multiple forms. First, even if Conscientiousness exists to the same degree in all cultures, the inclusion of questions with a WEIRD-bias, such as the voting-question, means the scale will be a more valid measurement of Conscientiousness in WEIRD populations, and will thus be more capable of capturing individual differences that has to do with that trait in those populations. Suppose, for example, that every other item works perfectly across cultures and has the same differences between two groups everywhere. Then, the cross-cultural variation in group differences in Conscientiousness will be entirely driven by the validity issue with that item – smaller differences between groups should then tend to be found where the item works less well as a measurement of the trait of interest. With decreasing validity in further items, such influences will get even stronger, and may cancel out effects in other directions on items that perhaps work bet-

ter across cultures. Thus, such methodological biases may favor the assessment of individual and group differences in WEIRD-populations, while missing other differences that may exist in non-WEIRD-populations.

Beyond such issues with the content of questions, the particular assessment method may also influence the validity of results in different countries. One relevant result is that Westerners appear to have a more individualist (compared to collectivist) views of the self, compared to people in other cultures (see e.g., Henrich et al., 2010b). Although this has primarily been assessed with a limited number of non-Western countries, primarily Japan, China, and South Korea, so the pattern across countries may be more complicated than simply categorizing the West from the non-West. Such a perception may make self-assessments in terms of personality characteristics (compared to e.g., social roles) more central for how people view themselves (see discussion in Costa et al., 2001). Thus, even if the content of measures is appropriate in all cultures, the way in which this content is measured may cause a bias towards Western cultures. Suppose, for example, that there is an increasing tendency for women to behave conscientiously in all cultures compared to men (the direction found in Mac Giolla & Kajonius, 2019). However, if people in less individualist countries do not explain this with reference to individual personality characteristics to the same degree, but rather for example in terms of their social roles, then the personality trait assessment will be less valid a measure of such conscientious tendencies there.

Additionally, individualist attitudes may be associated with an increased tendency to stand out (see Henrich et al., 2010b), indicated by how participants in Western countries appear more likely to answer with the more extreme response options in the scale compared to more collectivist cultures. That Asian countries had smaller variation in scores (which suggests answering with middle options) compared to Western countries, was for example found by McCrae (2002) and Schmitt et al. (2008). More extreme responses may increase the measured differences found between individual and groups. However, this would then be due to response styles as a function of such individualist tendencies, or tendencies to stand out, rather than other factors.

This issue with how validity problems may confound assessments at different levels may also be considered in relation to variation due to situations rather than a variation specifically due to cultures. These factors thus also relate to Paper I. When models are applied in one context, they may take into account the specifics of that situation to build the best model with the data available there, similarly to how measures developed in the West may be constructed and refined over time to work as best as possible within that cultural context. However, this does not mean that those measures or models will necessarily generalize to other contexts. In fact, if culturally- or situation-specific factors are central for constructing the best model/measure in that culture or situation, then this can be expected to worsen the behavior of that model/measure when applied in new contexts.

Summary and Applications

We have now gone through some challenges when trying to measure psychological constructs, that may limit models' generalizability, as well as the ability to compare results across contexts. Next, we turn some more detailed applications of these ideas relevant for the papers in this thesis. Two such applications are considered: machine learning methods and their adaption/vulnerability to a specific context, and the possible cultural biases of the gender-equality paradox.

Machine Learning Models of Personality

Here, we consider the importance of examining the generalizability of psychological models. Primarily, this will be applied to machine learning models, as used in Paper I, which considers such models for predicting personality. However, to the degree that a personality test can be considered a model of a certain personality trait developed in a specific context, it will also be of relevance when considering how such models generalize across cultural contexts, as considered primarily in Paper III.

Machine Learning Basics

Machine learning as a field is concerned with the training of statistical models to yield the best possible predictions. The difference to the broader field of statistics, if a difference exists, lies in their focus. Statistics as a field is often more focused on inference, such as hypothesis testing of theoretically derived and fitted statistical models, while machine learning, as noted, has as a primary focus to achieve maximal prediction, using general-purpose learning methods for extracting patterns in complex data (e.g., Bzdok et al., 2018).

When the goal is inference, researchers may be more constrained in the models they fit, as they are then based on theoretical considerations for which variables to include or not, and how they should be modelled more generally, to derive interpretable models for future theorizing. When the goal is to maximize prediction, however, such constraints may be more of a hindrance than as asset. Within machine learning, the goal of maximizing prediction is thus often accomplished with methods capable of capturing general patterns with multiple predictors, such as k-nearest neighbors, decision trees, and deep neural networks (see e.g., James et al., 2013, for detailed explanations of the concepts in this paragraph, although the details are not necessary to understand the concepts in this thesis). Such increased flexibility may risk overfitting to data, however. Thus, within machine learning, cross-validation is conducted to yield models with good predictive accuracy for new data. This means that a random subset of the data is put aside for each training-round. Different

models are then trained on the remaining data, after which their accuracy is checked on the data put aside. This may be repeated until all data has been set aside once, after which the accuracy is averaged across all hold-out samples. Which of the models that then has best such accuracy affects which model is kept as the final model (or for another round of cross-validation).

When cross-validation is conducted, this allows the tuning of hyper-parameters. As an example, k-nearest neighbors is a statistical model that predicts an outcome of a new observation with the average outcome from the k observations in the model closest to the new observation on a set of predictors (according to some measure of closeness). The value k (the hyper-parameter) has then been decided by the cross-validation in the model training stage, by minimizing the prediction error of the model when applied to new data. If, for example, $k = 3$ resulted in the best accuracy on new observations in the training stage, then the three closest observations are used in the model. See Figure 1 for an illustration of how model complexity tends to affect model error (see also e.g., James et al., 2013). The same general procedure can be used for deciding whether to include a predictor, or which method to use for prediction (e.g., k-nearest neighbor, decision trees etc.). At the final stage, accuracy of the final model can be assessed with a test sample that has been randomly set aside before the cross-validation procedure.

Thus, the focus in machine learning is on ensuring the best accuracy for new observations. This has been characterized as it being concerned with finding “generalizable predictive patterns” by, for example, Bzdok et al. (2018, p.233). However, by the terminology used in this thesis (and in psychology more broadly), their focus is on replicability, not generalizability. Indeed, the cross-validation procedure used in the training stage is perhaps the purest example of a replication: the sample is randomly split into one training and one test part (which should therefore function as two random samples from the same population), and the model, created in the training sample, is replicated and assessed in the test sample. This should thus maximize prediction for new observations from the same population and in the same situation in which the model was trained. However, this does not automatically mean the model will generalize well. In the next sections, we will therefore consider some aspects of model building (in machine learning and more generally) that may influence the degree of generalizability of such models.

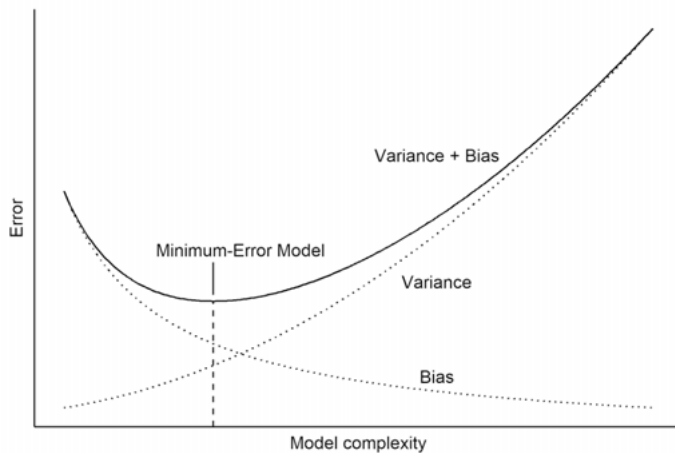


Figure 1: An illustration of how model error tends to depend on model complexity for a specific sample size and population (in machine learning and statistical modelling more broadly). Total error can be decomposed into bias (sometimes written bias^2), which captures systematic deviations between model and outcome (i.e., systematic error), and which tends to decrease with increasing model complexity. Additionally, there is variance, which captures non-systematic deviations (i.e., random error), and which tends to increase with model complexity (within machine learning, variance further tends to be decomposed into non-reducible error and [reducible] variance). The goal of cross-validation is to find the model with lowest total error in predicting new observations, which will mean finding a balance between bias and variance.

Model Misspecification

When statistical models are employed, such as exploratory factor analysis for identifying personality traits (e.g., Johnson & Wichern, 2013), the best-fitting model is found assuming that the model setup is correct. Consequently, if those assumptions are incorrect in some significant way, even models that replicate well across assessments may capture the psychological construct poorly, and lead to improper conclusions. As noted in the previous section, machine learning attempts to guard against this by using general-purpose models that can capture complex relationships. In machine learning terms, this means models with low bias but perhaps higher variance (James et al., 2013). However, complexities with model building can still lower the validity (increase the bias) of such models.

One general challenge in model specification comes from not including a more direct predictor in the model. Less direct predictors may compete about predictions, and win out in different situations, although the more direct predictor would tend to win out more generally and thus increase the robustness of the model. This can make models that have a high number of predictors, but lower theoretical grounding for those predictors, less robust (this idea will be examined in Paper I). Machine learning often include many predictors (see

e.g., Stachl et al., 2020; Youyou et al., 2015), so this may be a particular challenge to the generalizability of such models.

Another model specification issue may be challenging for any model of personality. Machine learning models that predict personality (and simpler prediction models) will set it up so the personality trait is treated as if it was dependent on the measured behavioral predictors (see e.g., Stachl et al., 2020; Youyou et al., 2015). However, from a biological entity perspective on personality, the behaviors depend upon the personality trait, not the other way around (McCrae & Costa, 2008). Even from an emergent perspective on personality (see e.g., Baumert et al., 2017; Cramer et al., 2012; see also van der Maas et al., 2006 for an application to intelligence), the personality trait is typically identified with the interrelated network between general (self-reported) personality tendencies, as such tendencies are typically what is assessed in personality questionnaires. However, when assessed with behaviors, only specific instances are used as predictors. A long enough measurement can perhaps average into predictors close enough to such overall tendencies, but barring that, the theoretical model would better fit the other way around: overall tendencies (which together make up the personality network, see e.g., Cramer et al., 2012) is what causes specific instances of behaviors. Consequently, as the personality modelling gets more involved, such misspecifications may limit the models' robustness, as it may not obey an enduring causal structure.

Another simpler but perhaps quite common model misspecification issue is to not include a relevant predictor that covaries with another included predictor. When that other predictor is not included, the one that is included may have either a stronger or weaker association with the dependent variable than it would otherwise. If both are positively associated with the dependent variable, and are themselves positively correlated, then the apparent association will be strengthened. This can factor into any correlational assessment, but is perhaps especially problematic in cross-cultural assessments where multiple cultural differences may covary. In gender equality paradox studies, as discussed in Paper II-III, this becomes a possible problem when cultural factors other than gender equality are not controlled for, such as the strength of gender stereotypes (see e.g., Breda et al. 2020; Miller et al., 2015). It also becomes an issue for machine learning methods when some relevant variables are not measured, say because they are harder to capture in a specific environment. This may in turn lead to other important measured predictors that interact with such predictors to not be included.

Model Complexity and Model Robustness

On the flip-side of model specification issues having to do with including too few relevant covariates and/or an incorrect model specification, large and flexible enough models may have their own issues. First, because they may be difficult to interpret, and thus provide less immediate understanding of the

examined relationship that can then be used for further theorizing. This is the balancing between prediction and explanation important for any modelling (see e.g., Möttus et al., 2020), including machine learning. More complicated models, such as neural networks trained on a massive set of predictors (e.g., Read et al., 2019), may provide superior prediction for any specific population and situation. Conversely, however, simpler models, such as linear regression using theoretically motivated predictors, may provide a more understandable model that may aid the goal of explanation.

As previously noted, increasingly complex models may also have too high variance (see Figure 1), leading to a deterioration in predictive power for new observations, as they become too affected by particulars of the training dataset (James et al., 2013). Such ideas about model complexity can further be associated with theories and results within cognitive psychology about the optimality of heuristics under high uncertainty (e.g., Gigerenzer & Brighton, 2009; Gigerenzer & Gaissmaster, 2011; Juslin et al., 2009). Such perspectives state that, when there is noise in the environment, that noise can limit how good predictions people can make from taking into account the full range of information, and when utilizing perhaps more accurate, but more complicated models. Instead, “heuristics”, or simplified prediction rules, which have a bias in what information they take into account, and/or how that information is weighted, can perform better overall. Even though that bias may cause them to break down in unusual but systematic situations. Consider the minimum error model in Figure 1 from before, which has some level of bias and variance. Optimal heuristics, then, are those which accept some amount of bias for a larger reduction in variance compared to a higher-variance lower-biased complex model.

When it comes to machine learning models, these ideas mean that the optimal model will find a balance between bias and variance when model complexity is decided. Thus, the most complicated model will be unlikely to win out. With an increasing number of observations, however, an increasingly complicated model may be found to be the most predictive when new observations from the same population and situation is sampled, as the predictions become more accurate (i.e., variance decreases). However, while such models may win out in predictions under direct replication, this is not guaranteed when the population or situation changes. This follows from several of the ideas we have discussed: If the model does not capture the causal structure, and does not include the most central predictors, then a model trained to have maximal prediction for a specific situation may be highly sensitive to particularities of that situation, and will break down when something in that situation changes. The balancing of bias and variance might illuminate this issue. As the situation changes, the bias of the model may increase, as the model was trained to predict in a specific situation that no longer holds true. However, variance will not decrease, as it still depends on the variation of the model in the situation when it was trained. This shows that total error can be expected

to increase with changing situations. However, it does not directly state whether a more or less complex model may hold up better. If the simpler model only includes predictors that hold in the situation on which it was trained, it will not generalize better than a more complex model that also includes predictors which hold between situations.

This illustrates the importance of theory when selecting predictors. If theoretical considerations can provide a guide to predictors that can be expected to hold up better across situations – and if that theory holds true – then a model trained on such predictors specifically may hold better across situations. This is examined in Paper I (here we select more psychologically meaningful predictors, but do not employ any more involved theory). That is to say, when models are constructed either for maximal prediction for a certain population and in a certain situation, or based on simpler models with more theoretically selected predictors, how well do those varying types of models then generalize. From the previous discussion, it seems a more involved predictive model would be more sensitive to the particulars of the situation the model is trained in. It should be noted, however, that this is neither an advantage nor disadvantage of such a model. Indeed, being capable of picking up predictors which only function in one situation, and increasing in predictive power in that situation, may be seen as a great advantage of more complex machine learning models. However, it does mean it is of interest to examine the degree to which different models are generalizable, to provide a greater understanding of the examined construct and its associated predictors for future examinations.

Limits and Opportunities of Machine Learning Models

In conclusion, machine learning models seem tailored to excel at the goal of prediction: trying to capture the studied concept as closely as possible, and maximizing the accuracy of predictions (for the studied population/situation). However, they may not be as suited for the goal of explanation: that is, identifying the causal underpinning and relationships between the studied concepts (that may generalize more readily across situations). Thus, it may be more difficult to build up theories of such psychological mechanisms from machine learning studies, especially from more complex and “black box” models. For this, inferential and less complex, but more theoretically meaningful tests and models of psychological constructs or mechanisms may remain of higher potential. As the building of theory about constructs and mechanisms is a primary goal of psychology, this type of modelling is thus likely to remain highly central. With this type of testing, it is possible to specify the model or function as completely as the theory specifies, which results can then be compared with. It is then also possible to construct models that take further dependencies into account, rather than merely predicting the final outcome as strongly as possible. It is, for example, possible to include mediating or moderating variables that may test potentially important causal pathways of interest to the

researcher. As a simple example of a theoretical model, it might be predicted that the willingness to pay for a lottery increases linearly with the expected payoff of that lottery, weighted by some unknown constants. The way that data compares with that model may then be fairly simply assessed, which allows an evaluation of whether the proposed mechanism fits peoples' behavior accurately or not. With more flexible modeling, as typically employed in machine learning, that flexibility may allow the same or similar patterns to be captured, and provide similar (or better) accuracy, but it may not as directly provide information about the functional relationship between the variables, and therefore, as discussed, be less informative for explaining the behavior.

With the goal of explanation, however, machine learning models may still function as a complement to such testing. Although it may be difficult to parse the importance of different predictors in the more complex of these models (e.g., neural networks), it is, for example, possible to restrict the predictors, yet use such more complex models to examine how high accuracy that can be achieved from different subsets of predictors. This should provide information about more or less important predictors that can then be used for future theorizing. Similarly, cross-validation and other machine learning techniques can be employed with simpler models, but with a greater range of predictors, that may then be examined for the strongest-, and more or less generalizable predictors across situations. The results of machine learning models may also be compared to those of more theoretically derived models. If machine learning models have significantly greater accuracy in a situation than the theoretical model, then this indicates that there may be something that model is missing, which may provide a cause for further examination. In conclusion then, while machine learning models may be less immediately tailored for the goal of explanation, their use in conjunction with classical methods may help further such goals as well.

Cross-Cultural Assessments of Psychological Gender Differences

The cross-cultural result that will be examined is the so-called gender-equality paradox: that countries with higher gender equality (as measured by e.g., equal pay, equal access to education etc.), have shown larger gender differences on several psychological scales. Gender differences have primarily been assessed with self-assessment measures (e.g., Falk & Hermle, 2018; Lee & Ashton, 2020; Mac Giolla & Kajonius, 2019; Schmitt et al., 2008; Schwartz & Rubel-Lifschitz, 2009), but also some other outcomes such as propensity for studying Science, Technology, Engineering, and Mathematics (STEM; Stoet & Geary, 2018), and playing chess (Vishkin, 2022).

Theoretical Background

The above results are considered paradoxical in relation to the prediction that gender differences could be expected to decrease as societies become more egalitarian and allow the same opportunities for men and women. A related prediction from social role theory is that “*variations in ecological, economic, and technological factors that influence the roles of men and women in society also should influence psychological sex differences relevant to those roles*” (Wood & Eagly, 2012, p.93). As roles get more similar, so are related psychological gender differences expected to decrease. Thus, from this perspective, gender differences should decrease to the degree that increasing (e.g., political and economic) gender equality leads to greater role similarity.

Conversely, it has been argued from an evolutionary perspective that as societies provide more opportunities for individuals to follow their preferences, innate biological differences between groups, such as men and women, will materialize more easily (e.g., Schmitt et al., 2008). Consequently, from this perspective, larger gender differences in more gender equal countries (that also tend to be higher in development level, wealth, democracy etc.), are believed to be the truest revelation of such differences, with a prediction that differences will continue to increase with greater such liberties. Gender equality has then been forwarded as one possible cause for such increasing differences, as it has been argued to similarly provide people with greater opportunity and liberty to express their innate gender-specific preferences. Several studies have been conducted informed by this hypothesis, and have often become highly cited (e.g., Schmitt et al., 2008; see also similar arguments in Falk & Hermle, 2018; Mac Giolla & Kajonius, 2019; Stoet & Geary, 2018). For example, in Schmitt et al. (2008), it is argued that “*sex differences in personality are vulnerable to restraining environmental pressures. As a society becomes more prosperous and more egalitarian, innate dispositional differences between men and women have more space to develop and the gap that exists between men and women in their personality traits becomes wider*” (p.179). In Falk and Hermle (2018) it is hypothesized that “*gender differences in preferences should manifest themselves only if both women and men obtain sufficient access to these resources to independently develop and express their intrinsic preferences*” (p.1). Similarly, Schwartz and Rubel-Lifschitz (2009) argue that “*increased gender equality should also permit both sexes to pursue more freely the values they inherently care about more*” (p.171).

More generally, however, it can be argued that even if greater opportunities to follow preferences is what leads to larger differences between groups on related measures, this does not specify why those preferences are different between groups. That is, if social expectations remain different for boys and girls across countries, which helps shape their behavioral preferences (see e.g., Charles & Bradley, 2009; Wood & Eagly, 2012), then this too may lead to larger differences in societies where those ensuing preferences have greater

opportunity to be expressed. It is, for example, acknowledged by Falk and Hermle (2018) that their results that larger gender differences in economic preferences are found in more egalitarian and wealthier countries (where such economic preferences have greater opportunity to be expressed) are also consistent with differences in social roles that continue to be expressed in such countries. Similarly, Charles and Bradley (2009) argue that a combination of remaining gender essentialist ideas, and individualistic “self-expressive value systems” may explain the increased gender segregation by field of study in Western countries, as people are influenced by such perceptions and norms around them when, for example, deciding on a field of study to specialize in.

In the section on the challenges with assessing causality, it was briefly noted how an emergent perspective on personality traits (e.g., Cramer et al., 2012) functions as an alternative to the biological entity perspective (e.g., McCrae & Costa, 2008). The biological entity perspective, which also suggests the universality of basic personality traits across cultures, has traditionally been the most prominent view of personality, and has been forwarded by central figures in the field (e.g., DeYoung, 2015; Eysenck & Prell, 1951; McCrae & Costa, 2008; Rothbart et al., 2000; see Möttus, 2016 for a review). As an example, in describing their theory of personality, McCrae and Costa (2008) write that “*Personality traits are a function of biology, and all human being share a common genome. Therefore, the structure of personality ought to be universal*” (p.169). Supporting this view, when English instruments have been translated to other languages, they usually show similar factor structures (see McCrae & Costa, 2008 for an overview). However, lexical studies that construct factors from the ground up in other languages have been less consistent (Saucier & Goldberg, 2001), and the personality items sometimes show less validity outside of the West (Laajaj et al., 2019). Although exporting Western measures has been most common, with a prediction of the universality of such measures, more lately, culturally specific examinations of personality have been increasingly advocated (e.g., Thalmayer et al., 2022). In contrast to the biological entity perspective, by an emergent perspective of personality, different, for example, extraverted behaviors covary because of their logical interrelations. That is “liking people” means one is more likely to “enjoy parties”, which in turn may further increase one’s liking of people (Cramer et al., 2012). This perspective can perhaps also be applied here in relation to the results on gender differences in personality across countries. In countries where people have greater opportunity to express (for example) different extraverted behaviors, such as going to different social outings, partying, performing, engaging in political or other types of debates et cetera (or conversely for more introverted behaviors), then differences in preferences for such behaviors across groups may feed into each other to a greater extent, and lead to increasing differences on extraversion as a whole.

Thus, measured gender differences may differ in magnitude, and be larger in Western countries, both for biological and social reasons (as well as for a

combination of these factors, and, as argued below, perhaps also for methodological reasons). One area where there is notable contrast between predictions, however, is for the gender equality measure. That is, if greater gender equality on its own unanimously led to larger differences, then the prediction that differences between men and women may decrease with changes in norms and expectations seem less credible. For this reason, we will examine to what degree this measure is associated with larger differences compared to some other relevant measures in Paper II-III.

In contrast to the above cross-cultural results, changes over time show more contradictory results, and seem to better support the prediction that differences may decrease. For example, the cross-country paradox found with the measures employed by Schwartz and Rubel-Lifschitz (2009), was not replicated temporally by Connolly et al. (2020), who did not find a link with the gender equality measure over time (and found a decrease with time, albeit not connected to the measure). Similarly, Stoet and Geary (2018), found a larger gender segregation by field of study in countries higher in gender equality. However, this contrasts with the increasing proportion of women who have entered into STEM-fields in the more gender equal of these countries over time (this is for example illustrated in the supplementary information in Paper III). Further, over time, women have increasingly entered paid employment, and high-status jobs, and have shown associated increases in various indicators of agency, and leadership, with resulting smaller gender differences in “culturally masculine” attributes, social behaviors and vocational interests, while similar changes in more culturally feminine attributes show less changes, explained by how men have not shown similar changes into traditionally feminine roles (see Wood & Eagly, 2012 for an overview). In Paper II, I also show that the difference in the names given to boys and girls is smaller towards the end than towards the beginning of the assessed time-period (however, as noted in connection to this paper, this assessment method may have other confounds).

In corresponding with an author of the article which Paper II comments upon (i.e., Vishkin et al., 2022), I was sent a list of papers argued to show the paradox over time. However, these assessed decreases in mental health, and appeared to involve other more relevant factors over time than gender equality. To illustrate this point, in Högberg et al. (2020) and Thorisdottir et al. (2017) adolescents showed larger gender differences in the end than in the beginning of their timelines (at most 24 years) on mental health measures, as girl’s self-reported mental health decreased over time (more-so than boy’s). However, their timelines both extended across 2008, the time of the worst financial crisis since the Great Depression. Thus, their results both seem in line with women reporting more stress and anxiety following this crisis. In Högberg et al. (2020), much of the increased difference was also connected to increasing school stressors over time. Additionally, these authors did not dis-

cuss their results in relation to gender-equality, and poorer mental health appears to fit poorly with the interpretation that gender equality increases men's and women's ability to follow gender-specific preferences.

One reason why factors other than gender equality may be more important for the magnitude of gender differences across countries is perhaps that, while the employed gender equality measures capture such things as equality in education and pay, they do not capture gender norms and perceptions that may continue to affect men and women differently in different countries (as discussed above). Indeed, gender equality, like gender differences, is multidimensional. Thus, although it has, for example, been asserted that gender equal countries are those that most “*promote girls’ and women’s engagement in STEM fields*” in connection to the study on gender differences in propensity for STEM (Stoet & Geary, 2018, p.591), this is not actually captured by the gender equality measure. Instead, it captures such things as the ratio of female to male labor force participation, and female to male gross tertiary enrollment ratio (see Boulicault, 2020, for further discussion). Consequently, there is a possible validity issue with this conceptualization of the measure: it is only valid if countries that are more equal on the above measures also are those countries which most promote women's engagement in STEM. However, as the examination is made cross-culturally, differences in perceptions about genders, as previously discussed, may also play into the results. That math-gender stereotypes indeed appear stronger in countries with higher gender equality, and where there is a larger gender gap in intention to study math-fields is supported by the results in Breda et al. (2020). Notably, such countries tend to be Western. A similar pattern for women's participation in science and national science-gender stereotypes were found by Miller et al. (2015). This might be associated with particular gendered perceptions about math and science in this region (see e.g., Gigerenzer, 2022, for a historical overview).

As noted in Costa et al. (2001), gender stereotypes on a set of adjective check lists had similarly previously been found to be more (measurably) differentiated in the Protestant West compared to in the Catholic West as well (Williams & Best, 1990). The authors argue this may be associated with a lessened status of women following the Protestant reformation. This, as women's role in religious life diminished with, for example, the closing of the female convents, lessened role of the Virgin Mary, and as religious leaders advocated a separation between private and public life, with women argued to belong to the former. Although these results might also be associated with measurement factors, as discussed below. Additionally, it is possible that gender roles and expectations change with increasing development and related factors. As an example, eating disorders have historically been most prevalent in the West, although over time they appear to have increased in other countries as well, such as in Asia. This has primarily been explained by an increasing exposing to Western body ideals, although, other factors, such as changes

in lifestyles and changing stressors have also been implicated (see Pike & Dunne, 2015, for a discussion).

Thus, as such norms and perceptions are not captured by the gender equality measures, they may continue to be influential. Perhaps there may even be a bias in selecting gender differences which correspond to gendered perceptions in the same world-region where the gender equality measures are higher (the West) and/or are more capable of finding differences there for cultural/methodological reasons. Such a possible bias may lead to this cross-cultural association, although it might not hold when other assessment methods are employed. In the next section, we therefore examine the possible influence of some such cultural/methodological factors.

Some Possible Cross-Cultural Confounds

The possible cultural bias in the measures of gender differences might be seen in how measures tend to also be developed or selected from the same region from which the studies were conducted (the West). Whether this is Five Factor measures developed in English languages (Costa et al., 2001; Mac Giolla & Kajonius 2019; McCrae, 2002; McCrae et al., 2005; Schmitt et al., 2008), measures such as STEM-propensity (Stoet & Geary, 2018) that appear associated with stereotypes about math (which tended to be, measurably, higher in Western countries; Breda et al., 2020), or participation in Western (rather than for example Chinese) chess (Vishkin, 2022). Thus, the cross-cultural variation in differences may be affected by how culturally different countries are to the Western, Educated, Industrialized, Rich, and Democratic (WEIRD; Henrich et al., 2010a) countries from which the studies are employed and in which the measures are constructed. Thus, when the West reveals the largest differences, it may be that differences are larger there because of its culturally particular perceptions about genders (or about individual differences more broadly), and because measures have been selected to assess those perceptions and in that cultural context specifically.

This may be an especial issue considering that cross-cultural assessments may have validity problems if one's measurements do not work as intended in certain countries, or if there are validity issues with comparisons across countries. One proposed factor which we will not be able to assess in this thesis, but nevertheless appears important to acknowledge, is the possible influence of self-comparison processes (Guimond et al., 2007; Heine et al., 2002). This means that when people rate themselves on some self-assessed trait, this rating may be influenced by the level of that trait of people around the respondent. That is, if everyone in a society has high/low levels of extraversion, someone average for each society may rate themselves towards the middle of the scale, which would complicate comparisons between those societies (Heine et al., 2002). For the gender difference studies, it has primarily been suggested that people in less gender equal societies are less likely to compare themselves

across genders, leading to smaller measured differences (Guimond et al., 2007). However, perhaps a similar pattern could also be found if men and women are in fact more similar in more gender equal countries, but there remain different norms for men and women that causes some peoples' answers to gravitate towards one end of the scale. Such gravitation may be more influential with smaller differences, as it is easier to consider oneself to stand out compared with average others in that case, possibly further influenced by a need for self enhancement, that appears more prevalent in Western cultures (Heine & Hanamura, 2007).

Further, relevant for the controls in Paper III in this thesis, when measurements developed and validated in the West are applied in other cultures, validity, as well as reliability, may deteriorate and complicate cross-cultural comparisons (the notion that personality assessments deteriorate outside of the West is suggested by, for instance, Laajaj et al., 2019). Supporting such methodological effects, before the introduction of the evolutionary interpretation of the gender-equality paradox, it was found that cross-cultural gender differences had a strong correlation with a combined indicator of data quality. In fact, that correlation is stronger than later correlations found between gender equality and gender differences ($r = .71$, McCrae et al., 2005, p.554; see gender equality correlations reported in Paper III for comparison).

As previously discussed, there are also cross-cultural differences in the degree to which people view themselves in terms of personality and other individual characteristics rather than perhaps in terms of social roles (see Heine, 2008; Henrich et al., 2010b for reviews), which is perhaps also associated with the degree to which they answer with more extreme response options, rather than suppressing such answers. This may cause personality measures to work better, and reveal larger differences in the West (personality and similar self-assessment measures are the most commonly employed in gender equality paradox studies). Already the first study that found the gender-equality paradoxical results in personality (Costa et al., 2001) discusses the possibility that the larger (gender) differences found in the West could be a result of such measures being more relevant for individualist Westerners in how they understand themselves. Similar to the results for data quality, these authors found a correlation of .71 between countries' measured gender differences, and countries' estimated level of individualism (p.329). Further, Western countries have been found to have larger variation in answers compared to Asian countries, which may cause them to have larger differences between groups as a result of an increased tendency to answer with extreme response options (see e.g., McCrae, 2002; Schmitt et al., 2008). In Paper III, we examine how some of these factors, that may explain the pattern in complementary ways, are associated with gender differences across countries.

Aims

From the previous discussions, there appear to be several complications that can sometimes make it difficult to generalize psychological measures and models across situations. Thus, the overarching aim of this thesis is to examine the generalizability of some such models, as well as to take into account additional contextual variables that may influence the results found across situations.

This examination is applied to two assessments that may be particularly sensitive to such situational factors, as presented in the introduction. First, it is examined to what degree different machine learning models of personality generalize across different situations. Next, the situational consideration is applied to a psychological interpretation that relies heavily on cross-cultural results, the idea that gender equality leads to increased psychological gender differences, and which may therefore be particularly sensitive to confounding differences across cultures.

In **Paper I**, we examine to what degree a text-based machine learning model based on a broad set of atheoretically selected predictors and constructed to maximize prediction within one of two specific situations (personal essays or Reddit messages), generalized to the other situation. This is compared to the generalizability of a less complex machine learning model based on a restricted set of more psychologically meaningful predictors.

Paper II is a commentary to a study which, instead of assessing the association between gender differences and equality across countries, does so across U.S. states, and over time within the U.S., while still finding the paradox. However, there are issues with the statistical methods used in this study, and there are also differences in the cultural background of people within countries. Thus, I examine whether the paradox holds up to differences in statistical methods and with a cultural control.

In **Paper III**, we re-examine the results of multiple cross-cultural studies on gender differences with country-level data available. We go through conflicting results, examine how results hold up to cultural and methodological controls, and study whether gender equality or other cultural differences better explain the variation in gender differences across countries.

Paper I

Background

The popularity of machine learning techniques for the assessment of personality has exploded in recent years (see e.g., Argamon et al., 2009; Arnoux et al., 2017; Azucar et al., 2018; Bai et al., 2013; Kalghatgi et al., 2015; Kosinski et al., 2013; Majumder et al., 2017; Stachl et al., 2020; Tandera et al., 2017; Tausczik & Pennebaker, 2010; Yarkoni, 2010). However, while such models are capable of reliable predictions in the situation in which they are trained, it has not been examined to what degree they are generalizable across different situations (see Bleidorn & Hopwood, 2019; Tay et al., 2020 for theoretical discussion). Indeed, to add to researchers' understanding of personality, and to allow personality models to be applied appropriately, it is necessary to develop an understanding of whether the model applies to a wide range of situations (e.g., texts on Reddit, personal essays, blogs, Twitter), just a particular situation (e.g., Reddit), or even just in the particular situation, for the particular population, and during the particular time-period when the model was trained (e.g., young people who posted a lot of Reddit in 2020). Examining all situations are necessarily impossible, but the degree of generalizability can be examined when varying some of these factors. To help further theory development from machine learning models of personality, it should prove helpful in examining such generalizability of models.

In this paper, we train models on a type of predictor that can be applied in multiple situations and with differing levels of complexity: text data. The goal is to examine to what degree two different types of model building generalizes across situations: atheoretical and high-dimensional machine learning modelling aimed at the maximization of prediction in a specific situation; or simpler linear modelling based on a limited number of more theoretically justified predictors. High-dimensional here refers to the larger number of parameters that are considered when training the model, here mainly as a result of the larger number of predictors that are considered. While the former type of model may provide superior accuracy within the same situation in which it was trained, if it picks up predictors or associations which are more specific to the examined situation, then its generalizability may suffer. However, for the smaller more theoretically justified model to generalize, this requires that the predictors are in fact more relevant for predicting the outcome of interest across situations. As a more theoretical selection of predictors will naturally be more limited in

scope, the high-dimensional model may nevertheless have a greater opportunity to find predictors which do generalize.

Method

Data was gathered from two domains: personal essays by 2,344 students who also answered a Big Five test of personality (collected by Pennebaker & King, 1999); and messages on Reddit.com from 1,200 users who had indicated their Big Five personality scores on specialized forums. Words in these text corpora were classified into 73 LIWC-lists, which aims to capture psychologically meaningful groups of words, such as words having to do with “anger” (see Pennebaker et al., 2015). The proportion of each text which belonged to each LIWC category was then used for training the personality models. Thus, this should be a more directly psychologically interpretable list of predictors.

In addition to the LIWC categories, for the high-complexity machine learning models, the 20,000 highest term frequency-inverse document frequency (tf-idf) words and word-pairs (“bigrams”), as well as the 20,000 highest tf-idf level 1-4 characters (parts of words) were identified in the texts and included as predictors. This (tf-idf) means variables which are more common in some users’ texts (high term frequency), while also more specific to those texts (high inverse document frequency; i.e., they are not common for all users’ texts). Thus, this list contains less direct psychological meaning, but it allows for a more extensive search of candidate important predictors within a domain.

Data was split into 80% training and 20% test. For the low-dimensional models, forward selection ($\alpha_{in} = .05$) and Least Absolute Shrinkage and Selection Operator (LASSO) regression using 10-fold cross-validation (to find the shrinkage factor that yielded best prediction on new data) was used when training models for the five Big Five personality models (forward selection was included for comparison to the LASSO model, as it might provide an even smaller set of predictors). For the high-dimensional models, Support Vector Regression (SVR) was used when training models. Variables were standardized within each domain. Models were first trained on the training data in one domain (Essays or Reddit), and were then applied to the test data either in the same domain, or the other. The correlation between the model and the level of the self-reported personality trait was used to assess model fit.

Results

The correlations for the models trained on the Essays and Reddit data are presented in Table 1.1-2, respectively. Correlations were small to moderate within domain, which is common for machine learning models of personality (Azucar et al., 2018). The high-dimensional models had superior predictive

accuracy compared to the low-dimensional models when trained and tested within the Essays domain, and similar accuracy as the low-dimensional models for the Reddit domain. However, when assessed across domains the opposite pattern emerged. The high-dimensional models had inferior predictive accuracy when trained on Essays and tested on Reddit (in particular, the LASSO model significantly predicted all traits, with only so one for the high-dimensional models), while all models had similar (poor) accuracy when trained on Reddit and tested on Essays. There may also have been some differences depending on the personality trait that was predicted. Openness was predicted most strongly by the machine learning models within domain both for Essays and Reddit, although the differences were small.

In addition, we examined the importance of predictors (by their beta-weights), and the number of LIWC-predictors that remained in models first trained in one domain and then retrained in the other with only those predictors which had survived the first step. The number of predictors which survived both steps, compared to the number of predictors in the first step, was taken as another indicator of the situational specificity of models. We did not conduct similar examinations for the high-dimensional models as, per discussion with the data-science collaborators, it would be complicated to program and identify such variation in those “black box” models, where the underlying associations are harder to extract. Several predictors survived for the LASSO models (about 60%), but fewer for the select models (30%), and quite few predictors survived in both directions. This indicates that the most important predictors that were picked up may have depended a lot on the context. Of those predictors which survived both steps, many seemed quite central for the trait of interest (see e.g., McCrae & Costa, 2008). For example, Extraversion was predicted by several indicators of excitability: writing more words relating to “drives” and fewer related to “tentativeness” and “negations”.

Table 1.1. *Correlations [95% CI] Between Self-Assessed and Computer Generated (Trained on **Essays Data**) Personality*

Domain/Model	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism	M_r (SD)
Within domain (Essay model on Essay data)						
LIWC-select	.22 [.13, .31]	.07 [-.02, .16]	.13 [.04, .21]	.11 [.02, .20]	.16 [.07, .25]	.14 (.06)
LIWC-lasso	.25 [.16, .33]	.10 [.01, .19]	.15 [.06, .24]	.13 [.04, .22]	.16 [.07, .24]	.16 (.06)
Machine learning	.43 [.35, .50]	.39 [.31, .46]	.40 [.32, .47]	.37 [.29, .45]	.31 [.22, .39]	.38 (.04)
Across domain (Essay model on Reddit data)						
LIWC-select	.12 [-.01, .24]	.15 [.03, .27]	.10 [-.03, .22]	.14 [.01, .26]	.19 [.06, .31]	.14 (.03)
LIWC-lasso	.13 [.00, .25]	.15 [.02, .27]	.16 [.03, .28]	.16 [.03, .28]	.19 [.07, .31]	.16 (.02)
Machine learning	.06 [-.07, .19]	-.01 [-.14, .12]	.21 [.08, .33]	-.01 [-.13, .12]	.08 [-.05, .21]	.07 (.09)

LIWC-select = regression model based on LIWC dictionaries with only significant features included in the model, LIWC-lasso = least absolute shrinkage and selection operator based on LIWC dictionaries. The model was trained on 80% and tested on 20% of the dataset. All correlations are based on 20% of the dataset. **Boldfaced** correlations are significant on level .05.

Table 1.2. *Correlations [95% CI] Between Self-Assessed and Computer Generated (Trained on **Reddit Data**) Personality*

Domain/Model	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism	M_r (SD)
Within domain (Reddit model on Reddit data)						
LIWC-select	.13 [.01, .26]	.12 [-.01, .24]	.19 [.06, .31]	.19 [.07, .31]	.07 [-.05, .20]	.14 (.05)
LIWC-lasso	.13 [.01, .26]	.12 [-.01, .24]	.23 [.11, .35]	.25 [.13, .37]	.17 [.04, .29]	.18 (.06)
Machine learning	.30 [.18, .41]	.14 [.01, .26]	.17 [.05, .29]	.24 [.12, .36]	.17 [.05, .29]	.20 (.07)
Across domain (Reddit model on Essays data)						
LIWC-select	.06 [-.03, .15]	.03 [-.06, .12]	-.00 [-.09, .09]	-.03 [-.12, .06]	.06 [-.04, .15]	.02 (.04)
LIWC-lasso	.10 [.01, .19]	.07 [-.02, .16]	.03 [-.07, .12]	.04 [-.05, .13]	.06 [-.03, .15]	.06 (.03)
Machine learning	.10 [.01, .19]	.13 [.04, .22]	.07 [-.02, .16]	.07 [-.02, .16]	-.05 [-.14, .04]	.06 (.07)

LIWC-select = regression model based on LIWC dictionaries with only significant features included in the model, LIWC-lasso = least absolute shrinkage and selection operator based on LIWC dictionaries. The model was trained on 80% and tested on 20% of the dataset. All correlations are based on 20% of the dataset. **Boldfaced** correlations are significant on level .05.

Conclusions

The results largely supported that the high-dimensional models were more dependent on the domain in which they were trained, having superior or similar prediction within domain, but inferior or similarly (poor) prediction across domains. However, the situational specificity of models appeared quite high overall, as the predictors that survived depended a lot on the specific direction in which models were trained. Such strong influence from the situation on the predictors that comes out significant supports that it is useful to combine information from different situations to better understand how personality is expressed. This, as those expressions may differ much across situations.

In addition, we found differences depending on domain. While high-dimensional models never had reliably significant correlations across domains (having 1 and 2 out of 5 significant correlations), the low-dimensional models had so when trained on Essays (5 out of 5 were significant for LASSO, and 3 of 5 for Select), but not on Reddit (0 and 1 out of 5). This might be for several reasons. As almost all significant models were found when trained on the Essays domain, and as the machine learning models had higher prediction when trained and tested within it, it seems there may be some real effects of domain. The Essays dataset was larger, with about twice the number of observations compared to the Reddit dataset. Thus, estimates may have been more reliable, and therefore also generalized better. In addition, the Essays may have provided data that was either more relevant for assessing personality, or of higher-quality more generally. The Essays dataset consisted of personal essays written by students (see Pennebaker & King, 1999). This may have been more conducive for participants to express their personality. In comparison, the Reddit data consisted of messages in discussion boards, focused on discussing whatever subject was at hand rather than at expressing their own thoughts, feelings, and behaviors. In addition, the Big Five ratings were calculated by researchers in the Essays dataset, while final scores were self-reported in the Reddit dataset, and thus relies on subjects accurately reporting those scores.

In summation, these results show that there is information in examining models' generalizability, both for lower-dimensional and higher-dimensional models. High-dimensional model's greater range of predictors, more general-purpose models, and higher or equal prediction within domain did not help such models to generalize between domains. Instead, when high-dimensional models had improved predictions, this increased prediction appeared specific to the domain in which it was trained. Thus, it remains useful for researchers who utilize either type of model to examine their generalizability. Additionally, examinations of different domains may provide information about the types of situations that are more or less conducive for building more generalizable models.

Paper II

Background

The gender-equality paradox is the cross-cultural result that countries with higher gender equality have shown larger gender differences on various psychological measures (e.g., Falk & Hermle, 2018; Lee & Ashton, 2020; Mac Giolla & Kajonius, 2019; Schmitt et al., 2008; Schwartz & Rubel-Lifschitz, 2009), and some behavioral outcomes (Stoet & Geary, 2018; Vishkin, 2022).

The paradox has primarily been assessed across countries (e.g., Falk & Hermle, 2018; Lee & Ashton, 2020; Mac Giolla & Kajonius, 2019; Schmitt et al., 2008; Schwartz & Rubel-Lifschitz, 2009; Stoet & Geary, 2018; Vishkin, 2022) in cross-sectional data. In a break with this tradition which avoids some of the cultural confounds found across countries, Vishkin et al. (2022) examine it across time in the U.S. and in England and Wales, as well as across U.S. states. Vishkin et al. (2022) use as a measure the gendered-ness of names, based on their first letter, which are classified into voiced letters, which are argued to sound harder and more stereotypically masculine, and unvoiced letters, which are argued to sound softer and more stereotypically feminine. The authors present earlier results showing that a higher proportion of male names are voiced, and that experimentally constructed such names are associated as predicted. However, there appears to be limited evidence that such comparisons can be used reliably for real-world names, see Additional Considerations.

In their analyses, Vishkin et al. (2022) found a decrease in the female proportion of voiced names over years (for both the U.S., and for England and Wales), and an increase for male names (in the U.S. specifically). Across U.S. states, they found that both men and women have an increasing proportion of voiced names in states with higher female leadership equality, but not broader gender equality. This increase was higher for men than for women, resulting in a larger gender difference. In the article, the authors report the change in the female proportion as non-significant.

However, this study is based on some problematic methods, which this paper comments upon. First, Vishkin et al. (2022) test the gender-equality paradox in a manner that breaks with how it has been tested previously. In previous studies, the gender-equality paradox is primarily considered supported when more gender-equal populations have larger (measured) gender differences in the (Western) gender-stereotypical direction. Indeed, this is noted by the first author of Vishkin et al. (2022) in a separate manuscript, when presenting the

opposite (non-paradoxical) prediction: “*if gender roles are culturally constructed, gender differences can be expected to be smaller or even reversed in societies with greater political and economic gender equality*” (Vishkin, 2022, p.276). That is: the paradoxical prediction is that differences should be larger in more gender-equal populations, and the gender with which the measure is more stereotypically associated should have higher scores than the other (see also e.g., Costa et al., 2001; Mac Giolla & Kajonius, 2019; Schwartz & Rubel-Lifschitz 2009; Stoet & Geary, 2018). Thus, it is not considered paradoxical, from the perspective that gender roles are culturally constructed, if differences are larger in the non-stereotypical direction. Indeed, the evolutionary account of the gender-equality paradox relies on this directionality (e.g., Schmitt et al., 2008), as this account considers gender roles (and associated norms) to be at least in part based on innate differences between men and women.

In contrast, in Vishkin et al. (2022), the authors do not quite make it clear what they consider to support versus oppose their hypothesis. Although they suggest that in more gender-equal societies “*voiced names [should be] given more often to males and unvoiced names given more often to females*” (p.491), this can be interpreted as being either a prediction about slopes or “compared to each other”. Indeed, the authors’ presentation of the gender-equality paradox refers to this as a result about larger differences (see p.490-491), and the authors state that their results are “*suggesting a real and nonartifactual gender-equality paradox*” (p.494). Additionally, the results also suggest an increasing difference in the (larger) U.S. dataset over time and across states (see their Figure 1A and Figure 2). These aspects suggest the result may be about larger differences. Contrastingly, however, it might also be a statement about slopes specifically, that is: that men’s proportion should be higher in more than in less gender-equal societies, with the opposite pattern for women, and with no consideration of whether differences between genders are larger or smaller in more gender equal societies. During communication with the first author of the manuscript in the review process of this paper, he suggested that the latter interpretation was correct.

However, not requiring a larger difference between men and women causes complications. If differences are smaller in more gender-equal societies, this can be predicted by non-paradoxical accounts. For example, if parents started to more often gender-neutrally select their favorites from the names around them, and then adapted those depending on the sex of their child (e.g., they like Carla, so they name their boy Carl), then the proportions would close in on each other. Thus, if women started out with a higher proportion of voiced names than men, then the authors’ might find their predicted results, but differences would be smaller, and this would be for a non-paradoxical reason. As such, it remains of interest to examine whether the original way of testing for the paradox reveals any paradoxical results.

Additionally, the alternative test requires further clarifications. That is: is it enough that one proportion goes in the predicted direction, or should both

do so? Also, should there be a significant difference between the change for men's and women's proportions? Based on the writing from the first author (review comment on my manuscript), he suggested that the directionality was more important than a difference, and that the purpose of testing for an interaction between gender and the indicator of gender equality (time and female leadership respectively), was to ascertain that the trend was different for men and women. This suggests two requirements for documenting support of a gender-equality paradox: (1) at least one proportion is associated with higher gender equality in the predicted direction (higher proportion for men, lower for women); and (2) there is a significant interaction between gender and the indicator of gender equality in the predicted direction (men's proportion increases more than women's, or women's decrease more than men's). If these results are fulfilled, do we then not need to consider the directionality of the other proportion? That is: can both proportions, for example, increase with higher gender equality, as long as men's increase more? This seems to not respect the predictions about directions.

If the trend for one gender goes in the wrong direction, this seems to at most provide partial proof of the paradox according to these predictions about slopes. Indeed, if directionality is not considered for both proportions, just that the right proportion changes more strongly in its predicted direction than the other, then these predictions seem to consider a paradox supported by all results that would be so under the predictions about larger differences with higher gender equality. Meanwhile, they might also do so if results led to smaller differences. However, in another part of the review of this manuscript, the first author also argued that the female proportion was not significantly higher in states with higher female leadership, and that a part of this comment which argued that both proportions were higher in these states was therefore incorrect. Thus, I also considered a third requirement that seemed to obey this evaluation: (3) the proportion should not change significantly in the wrong direction for either gender (lower for men, higher for women). Thus, I consider (1)-(3) when examining the results from these predictions about slopes, but not larger differences. It is important to note, however, that after re-analyses and controls, (1)-(2) do not hold either.

In addition to the assessment and test-related questions above, there are also questions about the validity of the methods employed in this study. The main problem had to do with the assessment of changes across time. Measures over time should have dependence between timepoints. However, this was not modeled. In addition, one factor that may affect the proportion of voiced names is the cultural background/native language of the parents. The authors acknowledge that proportions may vary greatly between languages, so it is interesting to control for this factor. During the reassessments, I also found some possible problems with the assessments in earlier years.

Method

In reexamining Vishkin et al. (2022) Study 1, which examined the paradox across time, I began by simulating Type I error-rates under strong dependence between years using the author's methodology, as an illustration of how dependence may dramatically increase error rates. Next, I employed time series analysis in the U.S. dataset, as this accounts for such dependence between years. This was done by fitting ARIMA-models to the proportions over time (a general type of time series models that can account for non-stationarity over years). The assessed period was between 1880 and 2018, a long time period where gender equality would have increased overall in the U.S. By the predictions in Vishkin et al. (2022), there should thus be a systematic trend in the paradoxical direction. I did not analyze the English and Welsh dataset as it only included 10 separate years (once every 10:th year), which is very low for time series, and will lead to large uncertainties in the model-estimation. If there is evidence of a systematic increase or decrease across time, then time series models are improved by adding a drift-term. The best-fitting time series models for the male-, female- and difference in proportions were estimated using AICc (AIC corrected for small samples), and the significance of the inclusions of a drift-term in the best and neighboring models was examined.

In reexamining Vishkin et al. (2022) Study 2, I first observed that the simple slope analyses conducted by the authors appeared to include some error. I therefore recalculated the simple slopes to assess whether the change in proportions matched the predictions about slopes. Next, I controlled for the proportion of foreign-born inhabitants in states, as the language/culture of parents should likely affect the names they give to their children, and as the proportion of (popular) voiced names may vary between cultures.

Results

By the simulations, the method employed in Vishkin et al. (2022) Study 1 could result in highly inflated error rates with non-stationary dependence between years. By plotting the U.S. and England and Wales proportions per year, it was further revealed that there was strong dependence (shown by how neighboring datapoints were close together compared to the variation over time), that the linear model employed in Vishkin et al. (2022) did not fit the changes, and that the gender difference was actually smaller in the end than in the beginning. Additionally, the sources of the data revealed that there could be a bias in the registration of U.S. names in earlier years, as the database was first focused on registering workers, and even then, excluded significant subgroups of that population. Time series analysis revealed no systematic tendency for the male proportion to increase, the female proportion to decrease, nor for there to be an increasing difference over the years (although it did

reveal non-stationarity over time, making it highly problematic to not model dependence). See Figure 2.1 for an illustration.

The reanalysis of the simple slopes over U.S. states revealed that both male and female proportions of voiced names were higher in states with higher female leadership, although the difference increased. Thus, the prediction about the direction of slopes was not fulfilled for the female proportion. Additionally, the proportion of foreign-born inhabitants predicted the gender differentiation in the proportion of voiced names more strongly than did female leadership. See Figure 2.2 for scatterplots. When controlling for the proportion of foreign-born inhabitants, female leadership had no power in explaining the differentiation (female leadership did not have a significant interaction with gender), while the proportion of foreign-born inhabitants did. In addition, female leadership, the measure which had predicted the differentiation, was more strongly correlated with the proportion of foreign-born inhabitants across states than was the broader gender equality measure (which did not significantly predict a differentiation).

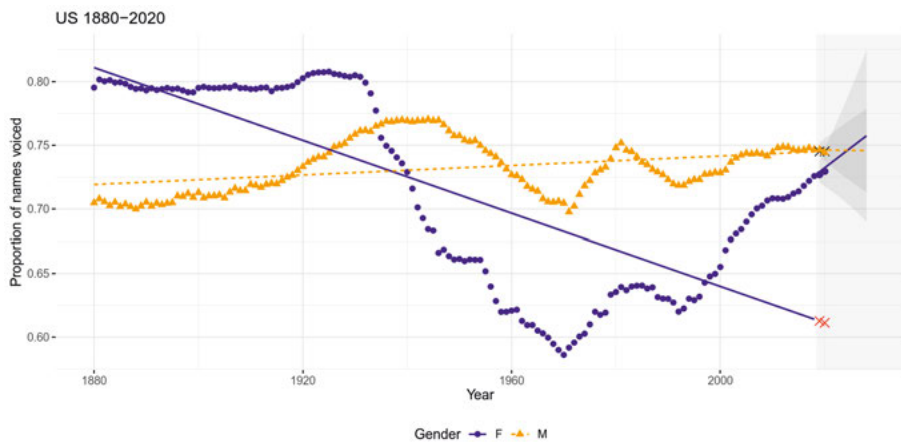


Figure 2.1: The proportion of female (blue dots) and male (orange triangles) given to newborns in the U.S. between 1880 and 2020 (as registered in the dataset), with multilevel regression lines for 1880-2018 from Vishkin et al. (2022). Grey background is for the years 2019 and beyond. Proportions for years 2019-2020 are plotted as for previous years. The authors' multilevel model's predictions for girls for those years are plotted as red crosses (far from the actual outcomes), and the predictions for boys are plotted as black crosses (and are more appropriately predicted). The prediction-line and confidence intervals (darker grey) resulting from the best fitting ARIMA models are shown for years 2019-2028.

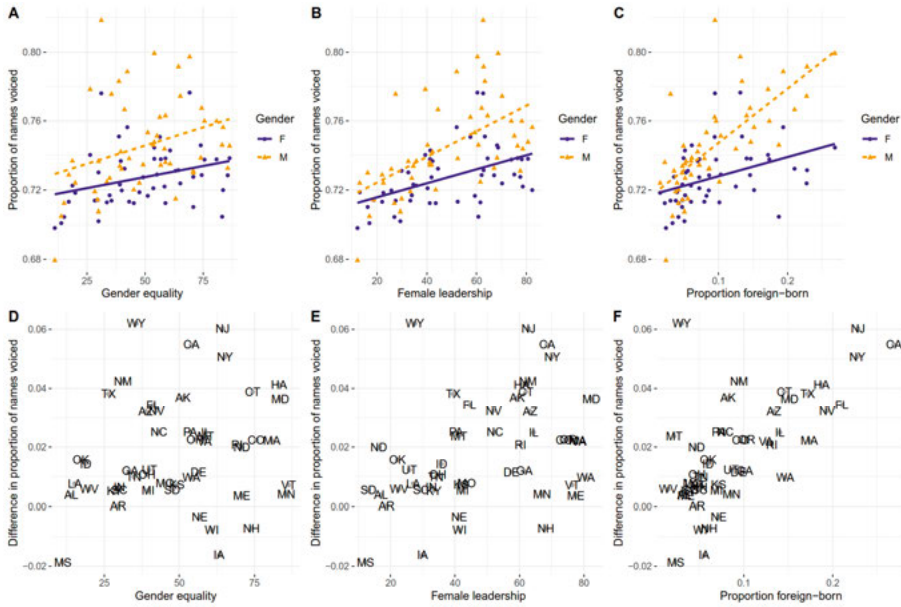


Figure 2.2: (Panels A, B, and C): Proportion of voiced female and male names given in the year 2018 in the U.S., with simple regression lines. (Panels D, E, and F): The difference in proportions (male minus female), with state-codes. X-axes show states' gender equality (A & D), female leadership (B & E), and proportion of foreign-born inhabitants (C & F).

Conclusions

The reanalysis of Study 1 revealed no trends in the directions proposed by Vishkin et al. (2022), nor a trend for an increasing difference. The female proportion, which changed the most, decreased most around the Great Depression and Second World War (turbulent times which, rather than gender equality, may have promoted more conservative values and stereotypes; see Thórisdóttir & Jost, 2011). This was further during a period when a larger part of the population may have become registered. Thus, at least part of the change may be due to a biased registering in earlier years. Indeed, as shown by the time series residuals in the supplementary information, the proportion for female names showed a change in behavior around this time point. After around 1970, the proportion has instead been increasing. Thus, there was no evidence of a gender-equality paradox over time with either method to test for it. More broadly, this is problematic for a causal interpretation of the gender-equality paradox on these outcomes, as gender equality must have had an effect over time if it is responsible for the variation in differences across, for example, countries or states.

Further challenging the importance of gender equality in explaining greater gender differences, there was no evidence that female leadership predicted any gender differentiation in proportions beyond the proportion of foreign-born inhabitants. In fact, the latter was a stronger predictor of the differentiation in proportions. Further, by predictions about slopes, there appears to only be partial evidence even without this control, as the female proportion of voiced names was higher in states with higher female leadership, contrary to the predicted direction.

In summation, there was no evidence of a gender-equality paradox with re-analyses that accounted for dependence between years, recalculated simple slopes across states, and controlled for cultural/language influences with the proportion of foreign-born inhabitants. This was true both by a prediction about slopes (no significant increase/decrease in proportions over years, no differentiation with controls), and larger differences (no larger difference over years, no differentiation with controls). In fact, in later years the difference in proportions was smaller, which may open up non-paradoxical accounts of the changes. However, it remains to be seen how the proportions change in the future, as the changes over time may just reflect random variation.

Additional Considerations

Here I present some discussion around the validation of the measure used. In addition, after publication of the article, some things have been pointed out that may require further clarification. For completeness, I therefore summarize some such additional considerations here.

Is the measure of gender differences well-validated?

First, it is worth considering the degree of validation of the gender difference measure, as it is interesting to consider how strong conclusions could have been drawn, had the results held to the re-analyses. The authors' rating of the masculinity/femininity of real-world names stems from their voicing, and then only the voicing of the first letter in the name. This follows a classification made by Slepian and Galinsky (2016). The idea that names beginning with a voiced letter are perceived as more masculine than names beginning with an unvoiced letter is indeed crucial for the conclusions drawn in Vishkin et al. (2022). However, the only referenced study which examines this perception for real world names is in Slepian and Galinsky (2016) Study 3, where 80 participants rated a total of 500 voiced names 0.175 points higher on average than 500 unvoiced names on a 7-point Likert scale, a rather small difference that may make it difficult to compare names across time and states only from their first letter. Further, those 1,000 names were treated as the sample, rather

than the 80 participants, so this may have overestimated the significance of results.

Indeed, the categorization using the first letter alone seems to introduce significant potential sources of error, as it seems it should be the name's sound as a whole which ultimately decides how masculine or feminine it sounds. In other studies (referenced by Vishkin et al., 2022) on the perceived masculinity/femininity and harshness/softness of names, this is controlled by having experimentally constructed names that control for everything except what is examined (Park et al., 2021; Pathak & Calvert, 2020; Pathak et al., 2020). Participants may for example rate "Fanatis" (unvoiced) and "Vanatis" (voiced) (Park et al., 2021). In such comparisons, everything except the first letter may be controlled for, and should therefore maximize its predictive power. However, real-world names will not have these controls. Thus, a name that begins with a voiced letter may for example be followed by mainly unvoiced letters, which should decrease the perceived masculinity of the name as a whole. Similarly, as related in Cai and Zhao (2019), other factors seem to influence the perceived masculinity/femininity of names as well: "*people tend to perceive a name as a female one if it is longer, has stress on a later syllable, or ends with a vowel rather than a consonant*" (p.63). Further, in the references Topolinski et al. (2014) and Topolinski and Boecker (2016), the aspect which affects their preference ratings of names is not decided by the first letter alone, but by how sounds move either from voiced to unvoiced (inward; front to back of mouth), or unvoiced to voiced (outward; back to front of mouth). The former is argued to elicit more positive emotions. As voiced names, by the classification in Vishkin et al. (2022) and Slepian and Galinsky (2016), all start with voiced letters, they will therefore contain all names that begin with an initial inward orientation. Thus, using the first letter as an indicator of masculinity appears highly confounded with this alternative explanation. Indeed, perhaps men are given more voiced names because they tend to sound softer/nicer as a whole (indicating, perhaps, a wish to move away from gender stereotypes), not because they sound more masculine. Further, in Topolinski and Boecker (2016) it is noted that "*the ultimate, not the initial, movement direction, determines affect*" (p.1590). If this is also true for judging the masculinity of names, or if initial and ultimate phonemes both affect it, this would make the initial phoneme measure as an indicator of perceived masculinity even more confounded.

Real-world names may also have associations that may be more important for how masculine/feminine they are perceived. For example, they may be associated with celebrities, religious figures, or extended family members. When parents select names for their children, those associations may be more important than the first letter. For example, a common name in the early data from the United States is Mary, which is classified as voiced, and therefore more masculine. However, it may also be influenced by the Virgin Mary, especially amongst the Catholic population. This association thus also appears

to include some stereotypically feminine norms that may have been more important for its selection. In summation then, there appears to be significant potential sources of bias in this measure.

How influential can bias in earlier years be?

As noted in the paper, there appears to be a biased registering in earlier documented years in the U.S. – as many as 40% of workers (and perhaps further non-workers) were excluded in the early registering. It is thus reasonable to consider how influential such a bias may be, especially as the unregistered population differed from the registered population in many important ways (unregistered people worked particular jobs, which for example led to more African Americans being unregistered in earlier years, and more generally appeared to exclude poorer working-class people). As pointed out after the publication of this article, if we treated 40% of women as unregistered and 60% as registered, and we treated the registered population as having 80% voiced names at its highest, and 55% at its lowest, then, the unregistered population would need to have only an 17.5% proportion of voiced names to explain the entire drop. Unless there are other biases in which names are registered, this thus seems unlikely. In the article, it is stated that a bias in registering may have caused the drop in the women's proportion around 1937. It is also noted that if some of the drop is real, it is also a special time-period that may be a more direct cause of the drop: it was in connection to the Great Depression and WWII, which may be a more direct factor. One possibility, as noted in the article, is because turbulent time-periods have been argued to lead to more conservative/traditional values (Thórisdóttir & Jost, 2011). However, it is also possible that the opposite is true: the time period may have led to less traditional values. Supporting such a possibility, the most popular name for girls in earlier years was the religiously associated name Mary (the Virgin Mary), which has then dropped in popularity. This is a voiced (more masculine sounding) name according to the classification in Vishkin et al. (2022). However, due to the religious association, it also appears to communicate some traditional gender norms for women. It is thus difficult to say whether the drop is because female names have been chosen to sound less masculine, or because such other factors (e.g., religious association) are of greater importance. However, the degree to which the drop around this time point may be due to a biased registering was thus not discussed in enough detail in the article.

Even if a bias does not explain the entire drop, however, it may still be influential for results over time. For example, if women's actual proportion of voiced names was around 70% instead of 80% in the earlier years, then their proportions increased more than they decreased over time, which could have influenced models over time. This would require a 55% proportion of voiced names in the unregistered population. If there was no problem with the registering, then we see a drop from around 80% to 55% in just 40 years in the full

population. Thus, it does not seem unreasonable that a non-random unregistered subset of the population would differ from the registered population to this degree. The male proportion is even more vulnerable to such differences. If the proportion was just 5% higher in earlier years than reported, then there would be practically no change in proportions in earlier compared to in later years. As the male proportion changed across these points in later years, it also seems like an unremarkable difference, that nevertheless could affect the trend over time. Additionally, as indicated by the smaller residuals in earlier years for the female proportions, the registering may have underreported the variability in the proportion of voiced names in earlier years. If so, then there may have been periods of some increase and decrease that now are unnoticed, which would make the consistent higher values for women or (slightly) lower values for men in earlier years less robust. All in all, then, there are many ways in which early bias can seep into results and alter conclusions.

Can one remove the names of children of foreign-born parents?

In the article, it is noted that female leadership is strongly correlated with the proportion of foreign-born inhabitants across states, and that female leadership cannot explain any differentiation in voiced names between genders beyond the proportion of foreign-born inhabitants (which is the stronger predictor, and explains beyond female leadership). Thus, as noted, female leadership is highly confounded with this measure. It would have been informative to identify which names belong to people with a different cultural background and control for this directly (e.g., by removing them). However, this data was, to the best of my knowledge, not available. Thus, I contended myself with noting that there is a confound.

Indeed, without data identifying names belonging to people with different cultural backgrounds, it is difficult to correctly control for this influence by the removal of names. The limited data existing which identifies names belonging to people with different cultural backgrounds seems unable to capture any significant proportion of this population. For example, in communication with the first author of Vishkin et al. (2022) after the online publication of this article, he suggested controls for names of children of foreign-born inhabitants should be done by removing people with names on https://www.baby-center.com/baby-names/most-popular/100-most-popular-hispanic-baby-names-of-2011_10363639 (checked November 28th, 2022) which lists the 100 most common Hispanic names in 2011. The author further argued that female leadership remained a significant predictor after these removals (although interactions in the authors' four main models, with proportions, all appeared weakened: one was non-significant already, and two had lower 95% CIs of 0.00). Hispanic and Latin American countries made up around 50% of foreign-born inhabitants in the U.S. in 2018 (<https://www.pewresearch.org/hispanic/2020/08/20/facts-on-u-s-immigrants/> accessed November 28th 2022).

However, this list is not the most common names in the population of foreign-born U.S. inhabitants, but the most common amongst names of newborns sent in to the Hispanic version of the babycenter-site (by users both within and outside the U.S.), for which the largest forum contains just over 11,000 users (active and inactive, U.S. and non-U.S. inhabitants), not all of whom would have had a newborn in the specific time-period (see <https://espanol.babycenter.com/c25001829/foros-buscando-nombres>, last checked December 12th 2022).

Further, even if these 100 names were the most common, they would only capture a proportion of all names of Hispanic and Latin American people. For example, the 100 most common male and female names in the U.S. in 2018 in the data for Vishkin et al. (2022) Study 2, only captured 33% and 41% of all female and male names given to newborns in that year. Thus, even assuming those proportions are a bit higher for children of Hispanic and Latin American foreign-born U.S. inhabitants, say 50% on average, and that the names from the babycenter-website captures the 100 most common names, then this would only remove names from 25% of all children of foreign-born inhabitants in the U.S. Considering that the babycenter-website is not an official government statistic for names in the full population, it likely captures an even smaller proportion. Additionally, such a control only focuses on a specific part of that foreign-born population, while for example Asian Americans are not accounted for. Finally, that list cannot separate children of foreign-born inhabitants from children of U.S.-born inhabitants. As it also contains multiple popular English names, controlling by removing people with names in this list would also remove a large number of people that we would not want to control for, and confound the control.

In summation, it would be informative to remove the names of children of foreign-born inhabitants in the controls specifically, to assess whether this is the factor that accounts for the variation. However, the possibility for such controls appears quite limited, as it cannot reliably capture most names of children to foreign-born U.S. inhabitants, and would be further confounded in which population it removes. In the article, I settled with controlling for the proportions of foreign-born inhabitants in each state, note that there is a confound, and that the predictions from female leadership could not be parsed from predictions from the proportion of foreign-born inhabitants. It is possible this is the factor that explains the association. However, it is also possible that some other co-varying factor explains the association even more directly.

Dependence over time in the English and Welsh dataset?

In the article, I did not analyze the English and Welsh dataset beyond mere description. Thus, one can ask whether this analysis really did have similar problems over time as did the U.S. dataset. It, for example, included names only each ten years. The short answer is that it should have similar problems.

The effect of years on the proportions was tested in Vishkin et al. (2022) without accounting for dependence between years. Thus, unless there is so little dependence between years that proportions can switch randomly between higher and lower values in just 10 years (something which neither the U.S. nor English and Welsh dataset seems to suggest), there will be dependence between years in the population (although this may be difficult to establish with only 10 time points in this data). In the Supplemental Information of the article, I noted that the proportion of voiced names for women in England and Wales between 1904-1994 appeared to show a similar pattern as the U.S. data (initial decrease followed by some increase). Although I did not include it in the final version of the paper, I made sure this held by examining the proportion of voiced names between 1996 and 2018 (<https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/adhocs/10429babynames1996to2018englandandwales>). For these years, all names with frequencies (for which there was at least 3 registered) were available, unlike only the 100 most popular names in ten-year intervals between 1904 and 1994, as examined in Vishkin et al. (2022) Study 1, so it should provide a more robust picture of the proportions. As illustrated in Figure 2.3, the female proportion continued to increase, showing that the previous decrease has not continued. Thus, the negative simple slope for women in England and Wales in Vishkin et al. (2022) Study 1 has similar problems as for the U.S. results.

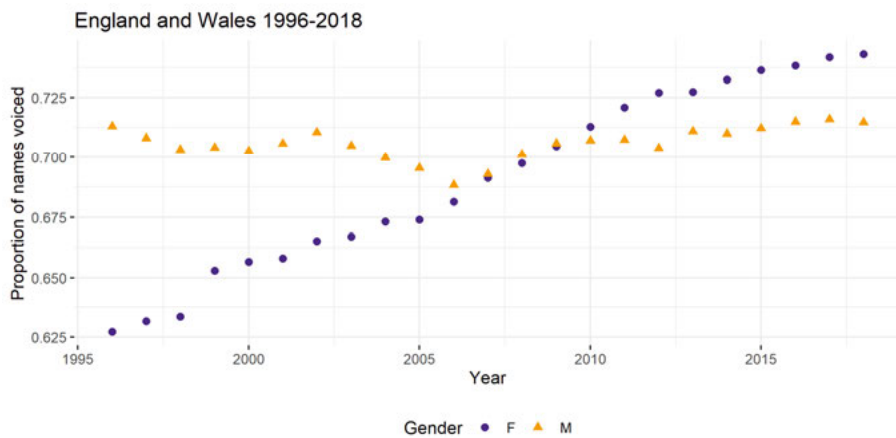


Figure 2.3: The proportion of female- (blue dots) and male voiced names (orange triangles) in England and Wales between 1996 and 2018.

Paper III

Background

With few exceptions (e.g., Vishkin et al., 2022), the gender-equality paradox – suggesting that gender differences are larger in more gender-equal societies – has been examined in correlational studies across countries. Interpretations differ between authors, however one prominent explanation for the correlation between equality and the size of gender differences is an evolutionary and causal interpretation. Increased gender equality is then believed to facilitate the expression of innate gender-specific preferences (see e.g., Schmitt et al., 2008; see also Falk & Hermle, 2018; Mac Giolla & Kajonius, 2019; Stoet & Geary, 2018; Schwartz & Rubel-Lifschitz, 2009).

A review of the measures used reveals a possible bias, however, such that the most commonly employed self-assessment measures were constructed to capture individual differences in Western countries. Indeed, no gender-equality paradox study examines the paradox by constructing culturally relevant measures of the most important gender differences that exist in, for example, Sub-Saharan African or Asian countries, and then examine whether those differences are larger in more gender-equal countries (i.e., Western countries). In fact, the perhaps most commonly employed assessment of such gender differences, on the Big Five personality dimensions, was constructed by first finding commonalities in English adjectives. This follows the lexical hypothesis, which states that important differences between people in a population will be encoded into the language of that population (e.g., Goldberg, 1990). Thus, if the most important differences vary by culture, then the Big Five should work better at separating individuals (and groups such as genders), in countries culturally similar to the English-speaking ones – perhaps in particular in other Germanic-speaking countries (e.g., English, German, Swedish). There are indeed indications of problems when assessing personality using translations of English instruments in studies on non-Western participants (e.g., Gurven et al., 2013; Laajaj et al., 2019; Saucier & Goldberg, 2001). In addition, gender stereotypes along dimensions mapped to the Big Five were previously found to be most (measurably) differentiated in Protestant countries (Williams & Best, 1990). Further, earlier studies found several indicators of poorer data quality in non-Western countries (Costa et al., 2001; McCrae, 2002; McCrae et al., 2005). With poorer data quality in some population, in-

cluding higher systematic or random error in responses, measurements become less capable of capturing the trait of interest, and tend to attenuate measured effect sizes (see e.g., Alwin & Krosnick, 1991; Curran, 2016; see also Loken & Gelman, 2017), such as differences between groups.

Even the measurements that are not made by self-assessment (and may therefore be less influenced by differences in psychometric properties across countries), appear to have a bias towards assessing Western perceptions about genders. The study by Stoet and Geary (2018) assesses propensity to study Science, Technology, Engineering, and Mathematics (STEM) in different countries. Intention to study math subjects has been found to be explained by the strength of stereotypes associating math with men in different countries (Breda et al., 2020), such stereotypes tending to be (measurably) stronger in the West (see also Miller et al., 2015 for results regarding science gender-stereotypes and the representation of women in science). Indeed, the West has a long history of gendering, for example, reason/logic (Lloyd, 2002), intelligence (Gigerenzer, 2022; Shields, 1975), science (Schiebinger, 1991), and math (see Hyde & Mertz, 2009). It is also the origin of much research on the idea of a separate female and male intelligence, that may therefore have had a particular cultural influence in this region (see e.g., references in Hyde & Mertz, 2009; Shiels, 1975, see also Gigerenzer, 2022). The study by Vishkin (2022) examines chess – also known as Western chess to distinguish it from similar games in Asia. As it has a longer history (with men) in the West, it should have stronger gender stereotypes there as well. It also appears connected to the gendering of logic and math discussed above (Lloyd, 2002; Gigerenzer, 2022; Hyde & Mertz, 2009).

There are also other potential issues with the assessment of the gender-equality paradox, stemming from the cross-cultural assessment. Indeed, gender differences and stereotypes were already found to be highest in the West, and in the Protestant West in particular (Costa et al., 2001; McCrae, 2002; McCrae et al., 2005; Williams & Best, 1990; see also Charles & Bradley, 2009 in connection with the STEM-study by Stoet & Geary, 2018), before the introduction of the causal evolutionary interpretation (Schmitt et al., 2008). Thus, the continued cross-cultural examination means that gender equality has not been tested as a causal antecedent of gender differences, but has merely been applied after the fact as a statistical covariate. Similar results would thus reasonably be found with other measures that are higher in the West (we illustrate this in this paper's supplementary information).

Speaking of causal interpretations, it is noteworthy that results over time (i.e., closer approximating the hypothesized process) are much less robust. Those that have found larger gender differences over time (albeit not predicted with changes in gender equality), are generally over a short timeline, meaning other factors may be more important. Specifically, in Högberg et al. (2020), and in Thorisdottir et al. (2017), gender differences in adolescent mental health were found to increase over time, as girls' mental health deteriorated

more notably. However, these trends were observed across the 2008 financial crisis, and, for Högberg et al. (2020), across a major school reform, which may be more direct causes. In contrast, there appears to only be one study that examines the paradox across time directly – by examining whether gender equality predicts changes in gender differences over time. In this study (Connolly et al, 2020; using the same measures as in Schwartz & Rubel-Lifschitz, 2009), the paradox was not found when gender differences were predicted with a gender equality measure over time, and instead they found a decreased gender difference with time (albeit not connected to the gender equality measure, possibly because gender equality and gender differences are assessed along different dimensions, see the Discussion in the paper). These results thus seem to contrast with the evolutionary interpretation that gender equality leads to increased differences. Further indicative of the difficulty with finding such patterns over time, women have received an increasing proportion of STEM-degrees in the West over a timeline of increasing gender equality, leading to a smaller difference in such degrees between genders (this is illustrated for U.S. and Swedish degrees in the supplementary information). Thus, the results by Stoet and Geary (2018) does not appear to be because these countries once had a larger proportion of women in STEM, which then dropped when their gender equality increased. Women have also increased in agency and aspiration into traditionally masculine fields over time more broadly, leading to various smaller gender differences (Wood & Eagly, 2012).

In sum, the cross-cultural gender-equality paradox seems to possibly have a cultural bias, as discussed in the introduction (Henrich et al., 2010a-b). Measures appear selected from a Western background with an aim to measure what is perceived as important (individual and) gender differences there. Thus, we examine if gender-equality holds any predictive power beyond cultural regions of various cultural distance to the West (and the Protestant and Germanic-speaking West in particular). Further, we examine whether a cultural difference account, as captured by a cross-cultural association between gender differences on the one hand, and cultural regions and data quality on the other (the latter of which should deteriorate with cultural distance as the measures become less culturally calibrated), better explain the variation in differences.

Method

We re-examined studies on the gender-equality paradox with country-level gender differences available either in the article, supplementary information, or over an internet link (Falk & Hermle, 2018; Lee & Ashton, 2020; Mac Giolla & Kajonius, 2019; Schmitt et al., 2008; Stoet & Geary, 2018; Vishkin, 2022) using the gender difference index of main focus on in these studies. We also examined the results with values in the European Social Survey (ESS) examined in Schwartz and Rubel-Lifschitz (2009) and Connolly et al. (2020),

although here we had to make some additional interpretations to, for example, calculate one overall measure of gender differences in countries (see paper). Gender equality was either gathered from tables in the articles, from the authors' sources, or in one case from a separate source. When multiple gender equality measures were included in the original article, we used the one that showed the strongest association with gender differences. In other words, we tried to be generous in choosing measures most likely to show the gender equality paradox, if there is one after controlling for confounds.

We then divided countries into seven broader cultural regions: the Protestant West (all countries that were either majority Protestant or majority Germanic-speaking), Catholic West, Latin America, Orthodox Europe, Muslim Middle East, Sub-Saharan Africa, and Asia. This follows common cultural classifications, while trying to only have regions with enough countries (for example, it closely follows the 2022 Inglehart-Wetzel cultural map of the world, except it has only one Asian region, and one Protestant region including English-speaking countries, while separating the large African-Islamic cluster into two regions based on their different histories, see <https://www.worldvaluessurvey.org/WVSEventsShow.jsp?ID=428>, retrieved December 20th 2022). We also calculated measures of data quality in the self-assessment studies (Falk & Hermle, 2018; Lee & Ashton, 2020; Mac Giolla & Kajonius, 2019; Schmitt et al., 2008, and the European Social Survey), primarily using measures of scale reliability such as Cronbach's alpha and item correlations. For Falk and Hermle (2018), when separating scale scores into individual items (see supplementary information for details), it was revealed that there were significant floor problems with certain items, so we also included the proportion of floor answers as a data quality measure in this study. If a scale is not wide or fine-grained enough to capture variation in some populations (or if participants otherwise always answer with the same option due to, for example, low motivation or understanding of the question), then it will be difficult to capture (gender) differences there.

For the analyses, we first examined baseline associations between gender differences and the three predictors used in this study: gender equality, cultural regions, and data quality, as well as for a simple indicator of Western cultural association (1 if Protestant West, Catholic West, or Latin America, 0 otherwise), to examine which better predicted gender differences. We also included some other variables higher in the West, including an Individualism estimate in the supplementary information. We then controlled for cultural regions and data quality to examine whether gender equality explained anything beyond these measures, and conversely, we examined whether they explained anything beyond gender equality. Controls for cultural regions were done with partial correlations (by mean-centering gender equality and gender differences within regions), and by entering regions and gender equality in the same regression model. This later method was also used when controlling for data

quality. If gender equality increased gender differences directly, then an association could be expected to be found also within cultural regions. To parse other cultural differences between regions from gender equality, the latter would need to explain something beyond such regions. For Vishkin (2022), who used weights for the total number of players in each country, we did the same by using weighted linear regression when mean-centering, then standardized the variables, and finally calculated a beta-weight for the controlled association in a weighted linear regression instead of a partial correlation. Finally, we examined the gender equality paradox in the Protestant/Germanic West in particular, as this is the region where the measures originate from, and where they should therefore have the highest relevance (they also tend to have large samples). Here we used partial correlations (as above) when controlling for the language of countries. We did this, as we observed that there appeared to be an opposite association in Mac Giolla and Kajonius (2019) for countries of the same language, compared to the world-wide results. This control may be especially relevant for the self-assessment studies, where differences in the associations of words and errors in translations may influence results.

Results

Our simple indicator of the West (1 if Protestant West, Catholic West, or Latin America, 0 otherwise) always had similar or (usually) stronger associations with gender differences compared to gender equality. The same was true for Individualism, and for other variables higher in the West. Cultural regions had additional explanatory power, and data quality was also almost always a stronger predictor than gender equality (the only exception was in Schmitt et al., 2008, which is the study with the fewest participants per country, which therefore may have noisiest estimates of gender differences and data quality). The data quality indicators sometimes dropped relatively little in magnitude, but the association remained strong. Table 3.1 shows explained variance between gender differences and the predictors. Gender equality never explained anything beyond cultural regions, except when the association reversed for chess participation (Vishkin, 2022). Controlling for data quality weakened the association in the self-assessment studies between 49% to 97% (average 77%), showing a strong confound, but for three of the five studies, the association remained significant ($ps > .02$). Figure 3.1 shows the confidence intervals for the correlations/beta-coefficients after controlling for cultural regions. Cultural regions always had additional explanatory power beyond gender equality, as did usually the data quality indicators (the only exception, again, was in Schmitt et al., 2008). Differences were most pronounced in Western regions (Protestant West, Catholic West, Latin America), and least pronounced in non- (or sometimes lately) Christianized regions (Muslim Middle East, Sub-Saharan Africa, Asia). This shows the confound between cultural

distance and gender equality. When entering cultural regions, data quality, and gender equality into the same model for the self-assessment studies, cultural regions were always significant, data quality was always so as well, except for in Schmitt et al. (2008), and gender equality was never significant. Additionally, when controlling for language regions in the Protestant West, the personality studies (Lee & Ashton, 2020; Mac Giolla & Kajonius, 2019; Schmitt et al., 2008) all revealed smaller gender differences with higher gender equality ($r_s = -.78$ to $-.90$, $p_s < .002$), with no other significant associations ($p_s > .15$). These negative partial correlations all remained significant with a combined control for data quality and language regions ($r_s = -.63$ to $-.90$, $p_s < .02$), although these exploratory results need further corroboration.

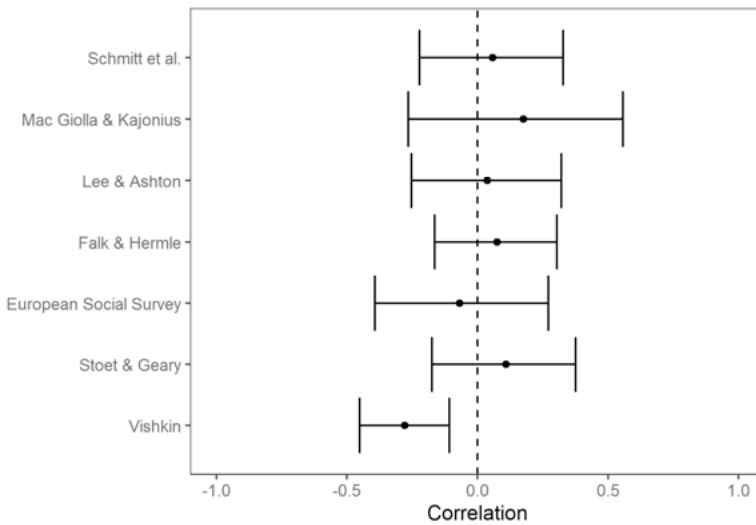


Figure 3.1: Confidence intervals for partial correlations/beta-coefficients between gender equality and gender differences in the reassessed studies when mean-centering within cultural regions.

Table 3.1.
 Baseline explained variance between gender differences on the one hand, and gender equality, the west indicator (1 if Protestant West, Catholic West, or Latin America, 0 otherwise), cultural regions, individualism, and data quality in each study. All associations were in the predicted direction.

Study	Gender equality	West Indicator	Cultural regions ^a	Individualism	Data quality
Schmitt et al. (2008)	.226*	.387*	.547*	.255*	.176*
Mac Giolla & Kajonius (2019)	.479*	.681*	.796*	.619*	.779*
Lee & Ashton (2020)	.352*	.523*	.753*	.372*	.494*
Falk & Hermle (2018)	.312*	.271*	.606*	.490*	.551*
ESS Values	.161 [†]	.397*	.427*	.235*	.238*
Stoet & Geary (2018)	.284*	.292*	.406*	.138*	--
Vishkin (2022) ^b	.145*	.214*	.418*	.258*	--

[†] $p = .0151$, $*p < .01$ (at least).

ESS = European Social Survey.

^aCultural regions with only one country are excluded when calculating explained variance. The average number of included countries per region were ~8 with Vishkin (2022) excluded (the study with the most countries), and ~10 with it included.

^bWeighted linear regression was used in accordance with the original method. For other studies, unweighted linear regression/correlation was used.

Conclusions

The gender-equality paradox studies appear to have a high potential for cultural confounds, and have possible indicators of methodological problems that may affect the pattern across countries. The association remained (albeit strongly attenuated) when controlling only for data quality. However, this is not a complete assessment of quality problems, so the association between gender differences and our data quality indicators may be indicative of further issues (see e.g., Curran, 2016), although this would require a more extensive analysis. Supporting this, floor effects in Falk and Hermle (2018) was also associated with differences in the predicted direction (larger floor effects, smaller differences). Additionally, for the values in the European social survey, as a check, we examined the association with gender differences for the number of excluded participants (due to not answering enough items, or answering too many items with the same option), and this too was significantly negatively associated. This is expected if countries with fewer careful responses showed smaller gender differences, and this control removed a subset, but not all, such responders. Further, when controlling for cultural regions (and language regions), the paradox disappeared or reversed. Cultural regions and data quality (when applicable) also appeared as stronger predictors of the paradox (further, Individualism and a simple Western indicator were stronger predictors than gender equality, as were other Western-associated variables). To parse the proposed effects of gender equality from those of other cultural differences, and of data quality (whether as a function of cultural distance, or because of other factors such as e.g., education level or quality of translations or interview situation), it seems necessary to assess the influence of these possible factors in greater detail, and to specify causal models that can be tested whether they correspond to the pattern of correlations found across countries. Further, other examination methods that assesses the causal claim without such cultural confound may be required. One such possibility may be to examine the effects of gender-equality reforms within natural experiments.

General Discussion

Summary of Key Results

Throughout three papers, it has been demonstrated how models and measures developed in a specific situation show several signs of deterioration when applied in a different situation. In Paper I, we found support that the degree of deterioration of models between situations depended on the complexity of the training of those models. More complex modelling that try to attain the strongest prediction in a specific situation may become more situationally dependent, compared to simpler models with a relatively more limited set of more theoretically based predictors. While it is intuitive that more complex models may lead to less reliable predictions due to overfitting, machine learning models guard against this with cross-validation for new observations from the same population. Generalizability, on the other hands, depends on to what degree patterns of associations remain similar across situations for the included predictors, which has been less examined in the literature on machine learning models of personality. It is, for example, not immediately certain that either type of model would have generalized, so by showing that the lower-dimensional models could do so while the higher-dimensional models could not, the results of this paper confirm the usefulness of examining the generalizability of different types of models, and of integrating this information for the building of psychological theories. In addition, our results suggested that generalizability was influenced by how the strongest predictors differed between domains to a large degree. Very few predictors survived in both directions when training in one domain after the other, even for the lower-dimensional models. Thus, situations may have a large influence on models of personality, and this might have been even stronger for the higher-dimensional models. Additionally, the results suggested that the domain that models are trained in may affect their generalizability, as some domains may provide better or more direct access to the inclusion of central predictors. It is possible that different behaviors of people in different situations (e.g., introspective or outward-looking), affects how well the model can be trained there. It may prove fruitful, for the understanding of how personality affects behavior, for future studies to examine whether certain domains reliably provide more generalizable predictors than others.

In Paper II, I found that when controlling for the proportion of foreign-born inhabitants, the gender-equality paradox across states disappeared. This may

illustrate how the measure used in this study can be difficult to use across different contexts, as differences in names across cultures can otherwise confound results. It is possible that this was a direct factor, but it may also be associated with additional confounds across states (for example, the U.S. born population can also differ in name traditions across states due to their immigration from various regions historically). Further, over time, the female proportion changed most during a period when a higher registering may have started taken place, which was also during the Great Depression and Second World War. This might also illustrate how unrecognized historical factors may also confound results. While the outcome measure may be noisy for the purposes used in the original article, one illustrative result of these re-analyses is that they demonstrate one example where a cross-cultural association (here over states), need not hold longitudinally over time. While higher female leadership (a measure of gender equality) was associated with larger gender differences in voiced names, this result was not found over a timeline of increasing gender equality. Instead, there was no systematic differentiation over time, and the difference was actually smaller towards the end of the timeline.

In Paper III, it was found that when controlling for cultural regions (and for language regions within the Germanic/Protestant West), the association between gender-equality and gender differences never remained, although sometimes it reversed so gender equality predicted smaller differences, challenging the robustness of the positive association. These results need to be replicated, however, and as with the world-wide results, other factors may be influential. Similarly, our data quality indicators (used in the self-assessment studies) always similarly or more strongly predicted gender differences than did gender equality. This is indeed what to expect if data quality deteriorated outside of the West, and if poorer data quality was at least a partial factor for the smaller differences in non-Western countries. Different types of data quality should be further considered in future studies, as well as different explanations for these results. Additionally, individualism, which have been associated with possible differences in how people relate to personality and other self-assessment scales (e.g., whether one answers more with middle options to the scale, and to what degree differences are explained in terms of internal differences or social roles), was a stronger predictor of the self-assessed gender differences. Finally, even the supplementary Western-associated predictors, that have not been forwarded as direct predictors of gender differences, typically showed similar or stronger associations. The associations with data quality and individualism are not new, but confirms previous results. Further, the various associated variables continue to illustrate the possibility for cultural confounds in the association between gender equality and gender differences across countries also for studies where these alternative accounts have not been assessed to the same degree.

We now turn to consider a couple of applications of these results.

Models may Interact with the Context

The broad point stated above is in some sense self-evident; different situations require different behaviors, so when we apply models that use those behaviors for some type of prediction, that prediction will likely depend on the situation examined. More interesting perhaps, is the degree to which such situational factors can influence models. In a recent paper, Yarkoni (2022) argued for a generalizability crisis in psychology. Here, the author illustrates the possibly large influence of contextual differences on results, by considering how unmeasured sources of variation, that the researcher nevertheless may intend to generalize their results over, can quickly increase the credible intervals of average effects to include zero. Thus, even fairly trivial variation may make small or average effects found in research difficult to generalize to other situations. Even for effects that are large on average, if their magnitude is also highly variable across situations, then predictions made from such an effect in one situation may not survive well when applied more broadly.

Machine learning models can typically achieve higher prediction than other simpler models in any given context. However, this does not mean that those predictions survive reliably. Instead, machine learning models typically need to be re-trained in another situation, or in all situations it will be applied to begin with. This illustrates machine learning models' tendency for general-purpose algorithms and large-variable datasets. This is not a deficiency of such models. Machine learning achievements can be impressive. As an example, recently much has been talked about a new Chatbot ChatGPT (see <https://en.wikipedia.org/wiki/ChatGPT>) that can produce detailed and convincingly human text. Such results show that machine learning can do very well in its ability to predict and emulate, for example, human language, if trained extensively enough (although it may still sometimes provide strange responses). Comparatively, machine learning models of personality as used in research have been far less extensive. It is thus possible that as training increases, and the training is conducted more broadly, such models will become capable of predicting more generally across the included, and perhaps similar, situations. However, to guarantee that this holds, such different situations need to be assessed in the first place. Further, although prediction may improve, it may also become difficult to assess what in the models it is that causes that improvement. That is, can the model predict personality because it has been able to find highly central and general predictors, or because it has managed to separate situations from each other, and include enough specific predictors for each situation?

Machine learning may have a larger part to play in future research. As Paper I indicates however, further consideration of the generalizability of such models, and a comparison between models of different levels of complexity, may provide additional understanding of what that added model complexity manages to- and does not manage to capture about psychological constructs

such as personality. Similarly, comparing high-dimensional machine learning models of, for example, personality, with more constrained and theoretically derived models, may also provide information about how much there may be left to predict that is not captured by such theoretical models. As discussed in the introduction, when it comes to the goal of explanation, of furthering psychological theory, machine learning may prove most useful when applied in such complementary ways to more theoretically derived tests and models.

Measures may be Cross-Culturally Confounded

The results of Paper II, and especially Paper III, also indicate a possible role of measurement confounds in the cross-cultural comparisons. In Paper II, the association between gender equality in female leadership and differentiation in the proportion of voiced names was confounded by the proportion of foreign-born inhabitants, and therefore perhaps by differences in language or culture. Differences in the frequency of phonemes, and how these tend to differ between genders, may thus make it difficult to compare this measure across populations. Further, in earlier years, women may have had a higher proportion of voiced names than men in the U.S. (if proportions are roughly correct), and this was also true (to a smaller extent) in earlier and later years in England and Wales. If so, it can be questioned whether an association of voiced names with men and unvoiced with women is all that general.

In Paper III, we demonstrated how the cross-cultural variation in the size of gender differences were highly associated with cultural regions. Western regions had the largest differences, while non-Christianized (or for some countries lately Christianized) regions tended to have the smallest differences. Although the association with regions (and some other confounds) has been noted in earlier studies, the magnitude of the association that can be associated with such regions (and confounds), and how other predictors are affected by such controls, has not been examined to the same systematic degree as done here, and functions as a reminder of the degree to which different cultural factors are intertwined. Further, gender equality did not predict anything beyond cultural regions unless the association reversed, while regions had additional explanatory power beyond gender equality. This shows that it is difficult to parse the proposed effects of gender equality for the larger differences in this region from other cultural idiosyncrasies in the West. Indeed, several previous results suggest there may be such confounds. In McCrae (2002) it is noted that Western countries had larger response-variances than Asian countries, suggesting perhaps an increased tendency to stand out rather than to fit in as a result of more individualist attitudes (see e.g., Henrich et al., 2010b), which would facilitate finding larger differences there. Further, our results about data quality indicate lower consistency in responses outside of the West, which may be associated with how the Big Five has shown less validity outside of

WEIRD-populations (e.g., Laajaj et al., 2019; Saucier & Goldberg, 2001). These results suggest it may be a better predictor of individual differences in the West, and perhaps especially in Germanic-speaking languages where it was first constructed (see e.g., Saucier & Goldberg, 2001), again facilitating finding larger differences there. Perceptions about genders may also differ between regions, and facilitate finding larger differences in the West when studies are informed by Western perceptions. As the gender equality measures used in these studies do not assess such differences in perceptions, they may miss accounting for such cultural differences. As noted, countries with larger gender gaps in intentions to study math subjects also appear to have more strongly gendered math stereotypes (Breda et al., 2020), with such countries tending to be Western (see also Miller et al., 2015 for results about science).

Several gender-equality paradox studies have found gender differences in the same direction across countries (although most included countries tend to be West or West-adjacent), with such differences largest in the West, and sometimes in particular the Protestant west. The interpretation has been that the largest differences in the West are therefore the truest revelation of innate such differences because they are more freely expressed there (e.g., Schmitt et al., 2008). However, as discussed, it is also possible that such differences are (measurably) largest in the West because of a combination of factors, such as measures being selected to capture variables of higher cultural importance there, and perhaps cultural perceptions about genders that are particularly prevalent there (see above). As an example, when it comes to the gender differences in personality (where men tended to show larger variation across countries), several differences appeared associated with an increased tendency to view oneself as socially independent/non-social, rather than as socially and emotionally connected with others (larger gender differences in extraversion, agreeableness, and emotionality, for which men scored lower than women; Lee & Ashton, 2020; Mac Giolla & Kajonius, 2019; Schmitt et al., 2008). In addition to the possibility that this is a truer revelation of innate differences, the difference may also be larger in the West because such social independence has been more of a norm for men in, for example, individualist societies, while simultaneous norms for women to, for example, be more communal and agreeable may cause less changes in the same direction, leading to larger differences. As the measures of gender equality do not rule out the existence of gender stereotypes or norms that affect the behavior of individuals also in these countries, it is indeed possible that gender differences on these dimensions are more expressed there, although still guided by such cultural perceptions. Additionally, differences in response styles (e.g., Guimond et al., 2007), or in the validity of measures in each region (e.g., Laajaj et al., 2019), may facilitate finding such increasing differences there.

In addition to gender differences' association with cultural regions, we also found consistent associations with indicators of data quality for the self-assessment studies. Data quality showed indications of deterioration as one

moved outside of the West (although the magnitude of this change varied). It is difficult to completely account for data quality, as there may be no single way to capture different types of lower-quality responses (see e.g., Curran, 2016). However, the findings suggest that at least some level of the increased differences found in Western countries may be attributable to such higher-quality data. Supporting this possibility, other data quality indicators, floor effects and the proportion excluded due to skipped- or too many similar answers, also predicted differences. There are several reasons why this association may come about. More educated and perhaps more well-off participants may be more capable and motivated to produce higher quality data. Additionally, there may be translation issues and differences in the quality of the assessment situation in non-Western countries. It is also possible that data quality differed because of lower cultural relevance or validity in populations, leading to less motivated participants, or participants who otherwise had less consistency in responses.

Such effects may result in confounding patterns when comparisons are made cross-culturally. The paradox is found in the direction where such cultural differences also appear to increase, and where data quality may decrease. Thus, future studies should further look into how such data quality issues may influence results, and additional studies should be conducted to find whether the pattern holds reliably without this cultural confound.

Limitations

In Paper II-III it was examined how gender equality compared with other candidate predictors of larger gender differences. This indicated that there were several possible confounding influences that might make (measured) differences larger in certain population (e.g., in the West in Paper III). This includes possible language factors (Paper II), and more methodological factors (in Paper III), whether this was an association with indicators of data quality, or factors that have been implicated in how people may answer self-assessment scales (individualism). This illustrates several possible confounds. However, it does not finally rule out that there are not also effects in the proposed direction for gender equality. So far, though, there appears to be little evidence for such effects over time, and some such results appear inconsistent with results across countries (STEM-students over time; Connolly et al., 2022; Wood & Eagly, 2012), nor has the association been replicated within cultural regions, or within countries. It is also possible that these other predictors themselves covary with some more direct predictors, rather than themselves explaining the association as directly. If there are theoretical accounts for such variables, attempts should be made to examine them more directly, and causal models for how different variables are interrelated should be specified, to allow them to be tested against the pattern of correlations found across countries. Further,

we employed the controls for cultural regions in Paper III partly because we expected this to parse other cultural confounds between such regions, while the continuous gender equality measure, by the causal hypothesis, should have an effect on differences both for countries within and between regions. However, regions are also associated with the magnitude of the gender equality measures. Thus, if regions are associated with several factors that causes measured differences to get larger in some of them, this may have swallowed up effects of gender equality (in whatever direction). Even so, however, this would still show that a simple bivariate model, where gender equality causes gender differences, does not sufficiently capture the associations found between countries. A direct effect of gender equality should be unlikely to be outcompeted by, and to be a weaker predictor, than less direct cultural predictors. The fact that the association disappears with cultural clusters suggests the baseline association is due to other confounds, alternatively that gender equality is a precursor to more immediate cultural factors. I am unaware of such a theory for cultural clusters, nor does this follow the evolutionary predictions, as they, as formulated, suggest a direct effect of the gender equality measures (see the Theoretical Background section of this thesis). As mentioned above, to be able to test different theoretical accounts, researchers should specify causal models for how gender equality influences gender differences in future studies. Finally, we found some negative associations with cultural controls, which further challenge a causal effect in the proposed positive direction. This was not true for all controls, however, and the ones that were found need to be further validated in future studies, especially the personality association in the Protestant West, which were found more exploratorily.

Additionally, the associations with data quality shows that data quality appears to decrease as one moves outside of the West, which is also where gender differences decrease. However, it does not show to what degree data quality is influential for such differences. Depending on the reason for why data quality deteriorated, it is possible that it explains more or less of the variation in differences. In addition to the association with our main data quality indicator, we also found some associations in the same direction with floor effects, and with the proportion of people excluded due to similar responding (another possible indicator of lower quality responses), suggesting some mechanisms in which data quality may have deteriorated in these studies. However, for most studies, we did not have other indicators than item consistency (Cronbach's alpha). In addition, it is possible that the influence of data quality differed between studies, so that although the association was in the same direction, the influence was stronger in some studies than in others. In future studies, different indicators of data quality may be examined more broadly.

When it comes to the machine learning models in Paper I, another possible limitation is that different models were not assessed beyond how predictors were included. It is therefore possible that another type of machine learning model (e.g., k-nearest neighbor) might have survived better across situations,

although this would need to be examined further. Similarly, although the low(er) dimensional models used more psychologically meaningful predictors, they were not directly trained according to some more theoretical account of personality. It remains to be explored whether theory-built models or high-dimensional machine learning models better generalize across situations, or if there is some changing middle ground where predictions generalize most readily (where to draw this line may depend on how well-understood the psychological construct is). Further, we only examined two different situations, and only received support for the generalizability of the lower-dimensional models one way. It is possible, as discussed in Paper I, that it was particularly conducive for finding more generalizable results, while many other situations might not allow equal generalization. It should also be noted that models were quite poor overall, judging by the strength of the correlations. The higher-dimensional machine learning models had similar prediction as has been found in other such examinations of personality from digital footprints (Azucar et al., 2018). However, this raises the question of how much prediction that can be achieved from such models, and how informative this is for the explanation of personality. Still, these models did achieve the highest prediction within domain, so this question may depend on to what degree it is possible to explain what it is in those models that provide this higher prediction. The easier it is to identify strong and central predictors, the more theoretical contribution should such models provide. As previously discussed, machine learning methods may be most informative when they are possible to use in tandem with more theory-based testing and modelling, such as when it is possible to parse information from such models to inform future theorizing. One way in which such models may be used in tandem might be to select different subsets of the predictors and examine how strong models than can be achieved with different subsets of predictors. Overall, the strongest result from Paper I might be that, even when high-dimensional machine learning models of personality can provide reliable and higher prediction within one situation, this is not a guarantee that such results will hold across situations. Our results indicate they may be even more vulnerable than low(er) dimensional models, but this should be further confirmed in future studies.

Future Directions

Future studies may further examine the generalizability of several results brought up here. When it comes to the machine learning models of personality, one informative procedure might be to examine how far different models can generalize, by examining multiple changes in the situation. For example, would the personality models trained on Reddit have survived better if they were first trained on certain forums, and then applied to others? How does this

change when Reddit is changed to another forum, or, as here, to personal essays? Different models will likely differ in how far such generalizations can be made. It may prove illuminating to examine whether certain types of model-building, with different predictors, reliably provide predictions in a greater set of situations than others. Such models or predictors may capture something more stable in personality expressions. Conversely, examining whether certain models or predictors reliably provide additional prediction within similar situations may allow a greater understanding of situational interactions with personality. Conjointly, examining to what degree it is possible to find such generalizable predictions, compared to more situational predictions, may provide additional information regarding the degree to which personality is expressed in similar or in different ways across different situations.

Regarding the cross-cultural studies, the results of this thesis suggest there may exist confounds due to cultural and methodological factors, and this may therefore require further corroboration beyond these cross-cultural comparisons. If it is possible to find natural experiments, for example with the introduction of gender equality reforms in some populations, then these may prove illuminating. This would also be a more direct test of the proposed causal effects. It may also be informative to further examine how strong and how central different perceptions or stereotypes about genders are in different countries, and how this is associated with gender differences found cross-culturally, as has been done with math and science (Breda et al., 2020; Miller et al., 2015). It may also be informative to examine how such perceptions change over time, and, for example, with increasing representation of genders in different domains. Additionally, as there may be cultural/methodological factors that influence the self-assessments, it could be examined whether the same pattern holds with behavioral outcomes. For example, do men and women also differ more in different behavioral expressions of extraversion (e.g., frequency of socializing, frequency/magnitude of positive affect) in the West compared to in other cultures? Similarly, it would be interesting to see how result would compare across countries when different populations are tested on instruments developed in each cultural context.

Finally, we have mainly discussed how methodological factors may have influenced the results. As discussed in the introduction, however, from a broader theoretical perspective, there are different explanation for results also beyond such methodological factors. It is possible that Western countries are indeed more conducive for larger differences on these variables, although the reason for that difference remains to be determined (e.g., innate biological differences, social norms and expectations, or a combination of factors). Conducting examinations like the ones above should help parse such different explanations. It remains to be explored to what degree such differences would change with changes in, for example, social norms, roles, and representation, and whether social factors can be influential enough to make them disappear or reverse, or to what degree differences remain. It is possible that this differs

depending on the measure. Perhaps average differences in personality, or in related developmental factors such as temperament, are indeed more stable than other factors such as values and choice of education. If this is the case, social norms or perceptions may still affect the size of differences on other outcomes, depending on to what degree norms about, for example, different fields of education, incorporate the possibility for such diversity. Further, although other factors appear more directly connected to larger gender differences cross-culturally, it is possible that increasing gender equality sometimes also has effects in the proposed direction, although this might also be for social reasons. For example, if stereotypically feminine traits have been less valued than stereotypically masculine ones in society, but women's political and economic representation increases, then this might empower the expression of such traits, which might lead to larger differences. However, if norms regarding these traits then become more inclusive both for men and women, the difference might decrease again. More research is needed to parse to what degree such differing factors are influential.

Conclusions

In this thesis, it has been examined how machine learning models and psychological measures that originate from a certain context can deteriorate or show different patterns outside of that context. This was true for how models of increasing complexity (and higher prediction within domain) showed less capability of generalizing across domains. As suggested by the results, situations can have a strong influence on which predictors come out strongest, and those strongest predictors need not hold in other situations. This was also true for how patterns of results may differ over time compared to across regions, and how comparisons of results across countries can be affected by cultural/methodological factors. As discussed in the beginning of this thesis the common thread has been in examining how such situational specificity, compared to greater universality, of psychological measures and models may affect comparisons across contexts, and require additional consideration. Such possible complications should not be a hindrance to a broader exploration of results, but rather prove the usefulness of attempts at validating results in varied ways. Whatever the effects of such increased attempts at generalization, whether results hold up across different assessment situations, whether they do not, or whether they provide further nuance to statistical patterns, exploring the generalizability of results should provide further information for psychological theorizing. The question of the generalizability of psychological results is far too broad to be fully addressed in this thesis. However, hopefully, the papers and discussion provided here may offer some part in that conversation.

Acknowledgements

First, I would like to thank my supervisors **Robin** and **Nazar** for your support and guidance throughout my work with this thesis and throughout my time as a PhD-student. I have benefitted greatly from your feedback, discussions, and encouragement, and I have truly enjoyed this time.

Håkan and **Karin**, thank you for your review of this thesis and for helpful discussions and comments during my final seminar that have truly helped improve the final version. Similarly, thanks to **Marcus** and **Ingrid** for your review and help with improving the articles presented for my half-time seminar.

To all my fellow, past and present, PhD-students, thanks for your companionship during studying and teaching, and for all the causal chats and socializing over coffee and outside of work. **Britt**, my first time as a PhD-student was made significantly more enjoyable by having you as a roommate at work, thanks for all the intriguing and creative talks, scientific and otherwise. Thank you also, **August**, for all your engagement and novel ideas. Thanks also to **Kim**, for organizing thoroughly enjoyable game nights; **Anton**, for our collaboration as co-chairs of the PhD association; and **Maja**, **Elina**, and **Olof**, for many hours of company during teaching.

To the members of the division for perception and cognitive psychology, **Peter**, **Ebba**, **Håkan**, **Leo**, **Ronald**, **Henry**, **Anders**, **Mona**, **Britt**, **Elina**, **August**, **Philip**, **Joakim**, **Mattias F**, and **Mattias L**, thank you for all the stimulating discussions during lab meetings, teaching, and in the corridors, and for so thoroughly including me in the group.

To my fellow RPG friends: **Britt**, for always making adventures more interesting; **Maja**, for your engagement with the world and its characters; **Martin** for keeping plans on track; **Olof F**, for always wanting to act heroic; and **Olof A-K**, for all the intrigues. Thanks for all the laughs, and for the opportunity to disconnect from work for a while.

Thanks also to my long-time non-RPG friends **Filip**, **Hannes**, and **Rosalie**, for your continual companionship during high school and beyond.

Finally, thanks to my parents, brother, grandparents, aunts, their spouses, and my cousins, for all your support. Mom and dad, **Jaana** and **Håkan**, brother **Jonathan**, and grandmother **Asta**, thanks for always being there for me, and for your belief in me.

To everyone mentioned and otherwise, to all colleagues, family, and friends, thanks for all your combined companionship and support. The strength it has provided has made this thesis possible.

References

- Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods & Research, 20*(1), 139-181.
- Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM, 52*(2), 119-123.
- Arnoux, P. H., Xu, A., Boyette, N., Mahmud, J., Akkiraju, R., & Sinha, V. (2017, May). 25 tweets to know you: A new model to predict personality with social media. In *Eleventh international AAAI conference on web and social media*.
- Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, Marcel A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*(2), 108-119.
- Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences, 124*, 150-159.
- Bai, S., Hao, B., Li, A., Yuan, S., Gao, R., & Zhu, T. (2013, November). Predicting big five personality traits of microblog users. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* (Vol. 1, pp. 501-508). IEEE.
- Baumert, A., Schmitt, M., Perugini, M., Johnson, W., Blum, G., Borkenau, P., Constantini, G., Denissen, J. J. A., Fleeson, W., Grafton, B., Jayawickreme, E., Kurzius, E., Macleod, C., Miller, L. C., Read, S. J., Roberts, B., Robinson, M. D., Wood, D., & Wrzus, C. (2017). Integrating personality structure, personality process, and personality development. *European Journal of Personality, 31*(5), 503-528.
- Berggren, M. (2020). *Conditional mean variables: A method for estimating latent linear relationships with discretized observations*. Unpublished Masters Thesis, Department of Mathematics, Uppsala University, Uppsala, Sweden.
- Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review, 23*(2), 190-203.
- Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203-219.
- Boulicault, M. (2020). *Measuring Gender Equality*. GenderSci Blog. Retrieved October 16th, 2022. <https://www.genderscilab.org/blog/measuring-gender-equality-why-the-gggi-is-not-the-right-measure-for-gender-equality-paradox-research>
- Breda, T., Jouini, E., Napp, C., & Thebault, G. (2020). Gender stereotypes can explain the gender-equality paradox. *Proceedings of the National Academy of Sciences, 117*(49), 31063-31069.

- Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233-234.
- Cai, Z. G., & Zhao, N. (2019). The sound of gender: Inferring the gender of names in a foreign language. *Journal of Cultural Cognitive Science*, *3*(1), 63-73.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81-105.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.
- Charles, M., & Bradley, K. (2009). Indulging our gendered selves? Sex segregation by field of study in 44 countries. *American Journal of Sociology*, *114*(4), 924-976.
- Connolly, F., Goossen, M., & Hjern, M. (2020). Does gender equality cause gender differences in values? Reassessing the gender-equality-personality paradox. *Sex Roles*, *83*, 101-113.
- Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R*. Odessa, FL: Psychological assessment resources.
- Costa P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: robust and surprising findings. *Journal of Personality and Social Psychology*, *81*(2), 322-331.
- Cramer, A. O., Van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., Kendler, K. S., & Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, *26*(4), 414-431.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.
- Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4-19.
- DeYoung, C. G. (2015). Cybernetic Big Five theory. *Journal of Research in Personality*, *56*, 33-58.
- Eysenck, H. J., & Prell, D. B. (1951). The inheritance of neuroticism: an experimental study. *Journal of Mental Science*, *97*(408), 441-465.
- Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, *66*, 68-80.
- Falk, A., & Hermle, J. (2018). Relationship of gender differences in preferences to economic development and gender equality. *Science*, *362*(6412), eaas9899.
- Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Department of Statistics, Columbia University, 348, 1-17.
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587-606.
- Gigerenzer, G. (2022). The idea of a peculiarly female intelligence: A brief history of bias masked as science. In *Intelligence in Context* (pp. 93-120). Palgrave Macmillan, Cham.
- Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107-143.
- Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*(1), 451-482.

- Goldberg, L. R. (1990). An alternative” description of personality”: the big-five factor structure. *Journal of Personality and Social Psychology*, 59(6), 1216-1229.
- Guimond, S., Branscombe, N.R., Brunot, S., Buunk, A.P., Chatard, A., Désert, M., Garcia, D.M., Haque, S., Martinot, D., & Yzerbyt, V. (2007). Culture, gender, and the self: Variations and impact of social comparison processes. *Journal of Personality and Social Psychology*, 92(6), 1118-1134.
- Gurven, M., Von Rueden, C., Massenkoff, M., Kaplan, H., & Lero Vie, M. (2013). How universal is the Big Five? Testing the five-factor model of personality variation among forager–farmers in the Bolivian Amazon. *Journal of Personality and Social Psychology*, 104(2), 354-370.
- Heine, S. J., & Hamamura, T. (2007). In search of East Asian self-enhancement. *Personality and Social Psychology Review*, 11(1), 4-27.
- Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What’s wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology*, 82(6), 903-918.
- Heine, S. J. (2008). *Cultural psychology*. New York: W. W. Norton.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). Most people are not WEIRD. *Nature*, 466(7302), 29.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). The weirdest people in the world?. *Behavioral and brain sciences*, 33(2-3), 61-83.
- Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences*, 106(22), 8801-8807.
- Högberg, B., Strandh, M., & Hagquist, C. (2020). Gender and secular trends in adolescent mental health over 24 years—the role of school-related stress. *Social Science & Medicine*, 250, 112890.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.
- Johnson, R., & Wichern, D. (2013). *Applied multivariate statistical analysis: Pearson new international edition*. Pearson Education Limited.
- Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, 116(4), 856-874.
- Kalghatgi, M. P., Ramannavar, M., & Sidnal, N. S. (2015). A neural network approach to personality prediction based on the big-five model. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, 2(8), 56-63.
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15), 5802-5805.
- Laajaj, R., Macours, K., Pinzon Hernandez, D.A., Arias, O., Gosling, S.D., Potter, J., Rubio-Cortina, M. & Vakis, R. (2019). Challenges to capture the big five personality traits in non-WEIRD populations. *Science Advances*, 5(7), eaaw5226.
- Lee, K., & Ashton, M.C. (2020). Sex differences in HEXACO personality characteristics across countries and ethnicities. *Journal of Personality*, 88(6), 1075-1090.
- Lloyd, G. (2002). *The man of reason: “Male” and “female” in western philosophy*. Routledge.
- Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, 355(6325), 584-585.
- Van der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological review*, 113(4), 842-861.

- MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, *55*(2), 95-107.
- Mac Giolla, E., & Kajonius, P. J. (2019). Sex differences in personality are larger in gender equal countries: Replicating and extending a surprising finding. *International Journal of Psychology*, *54*(6), 705-711.
- MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, *58*, 593.
- Majumder, N., Poria, S., Gelbukh, A., & Cambria, E. (2017). Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, *32*(2), 74-79.
- Mayo, D. G. (1991). Novel evidence and severe tests. *Philosophy of Science*, *58*(4), 523-552.
- McCrae, R. R. (2002). NEO-PI-R data from 36 cultures: Further intercultural comparisons. In R. R. McCrae & J. Allik (Eds.), *The five-factor model of personality across cultures* (pp. 105–125). New York: Kluwer Academic/Plenum Publishers.
- McCrae, R. R., & Costa, P. T., Jr. (2008). The five-factor theory of personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 159–181). The Guilford Press.
- McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*(1), 28-50.
- McCrae, R. R., Terracciano, A., & 78 Members of the Personality Profiles of Cultures Project (2005). Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology*, *88*, 547–561.
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437-455.
- Miller, D.I., Eagly, A.H., & Linn, M.C. (2015). Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology*, *107*(3), 631-644.
- Möttus, R. (2016). Towards more rigorous personality trait-outcome research. *European Journal of Personality*, *30*, 292-303.
- Möttus, R., Wood, D., Condon, D. M., Back, M. D., Baumert, A., Costantini, G., Epskamp, S., Greiff, S., Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A. G. C., Yarkoni, T., Ziegler, M., & Zimmermann, J. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the big few traits. *European Journal of Personality*, *34*(6), 1175–1201.
- Nilsen, E. B., Bowler, D. E., & Linnell, J. D. (2020). Exploratory and confirmatory research in the open science era. *Journal of Applied Ecology*, *57*(4), 842-847.
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600-2606.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.
- Park, J., Motoki, K., Pathak, A., & Spence, C. (2021). A sound brand name: The role of voiced consonants in pharmaceutical branding. *Food Quality and Preference*, *90*, 104104.
- Pathak, A., & Calvert, G. A. (2020). Sounds sweet, sounds bitter: How the presence of certain sounds in a brand name can alter expectations about the product's taste. *Food Quality and Preference*, *83*, 1–10.

- Pathak, A., Calvert, G. A., & Lim, L. K. S. (2020). Harsh voices, sound branding: How voiced consonants in a brand's name can alter its perceived attributes. *Psychology and Marketing*, 37(6), 837-847.
- Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX: Pennebaker Conglomerates (www.LIWC.net)
- Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6), 1296-1312.
- Pike, K. M., & Dunne, P. E. (2015). The rise of eating disorders in Asia: a review. *Journal of Eating Disorders*, 3(1), 1-14.
- Rantanen, J., Metsäpelto, R. L., Feldt, T., Pulkkinen, L. E. A., & Kokko, K. (2007). Long-term stability in the Big Five personality traits in adulthood. *Scandinavian Journal of Psychology*, 48(6), 511-518.
- Read, S. J., Droutman, V., Smith, B. J., & Miller, L. C. (2019). Using neural networks as models of personality process: A tutorial. *Personality and Individual Differences*, 136, 52-67.
- Rothbart, M. K., Ahadi, S. A., & Evans, D. E. (2000). Temperament and personality: origins and outcomes. *Journal of Personality and Social Psychology*, 78(1), 122-135.
- Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin*, 28(12), 1629-1646.
- Sackett, P. R., Lievens, F., Berry, C. M., & Landers, R. N. (2007). A cautionary note on the effects of range restriction on predictor intercorrelations. *Journal of Applied Psychology*, 92(2), 538-544.
- Saucier, G., & Goldberg, L. R. (2001). Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of Personality*, 69(6), 847-879.
- Schiebinger, L. (1991). *The mind has no sex?: Women in the origins of modern science*. Harvard University Press.
- Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods*, 8(2), 206-224.
- Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology*, 94(1), 168-182.
- Schmittmann, V. D., Cramer, A. O., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, 31(1), 43-53.
- Schwartz, S. H., & Rubel-Lifschitz, T. (2009). Cross-national variation in the size of sex differences in values: effects of gender equality. *Journal of Personality and Social Psychology*, 97(1), 171-185.
- Shields, S. (1975). Functionalism, Darwinism, and the psychology of women. *American Psychologist*, 30(7), 739-754.
- Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, 69(1), 487-510.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107-120.

- Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366.
- Slaney, K. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. Springer.
- Slepian, M.L., & Galinsky, A.D. (2016). The voiced pronunciation of initial phonemes predicts the gender of names. *Journal of Personality and Social Psychology*, *110*(4), 509-527.
- Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Völkel, S. T., Schuwerk, T., Oldemeier, M., Ullman, T., Hussmann, H., Bischl, B., & Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, *117*(30), 17680-17687.
- Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science*, *29*(4), 581-593.
- Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99-103.
- Tandera, T., Suhartono, D., Wongso, R., & Prasetyo, Y. L. (2017). Personality prediction system from facebook users. *Procedia computer science*, *116*, 604-611.
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24-54.
- Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining. *European Journal of Personality*, *34*(5), 826-844.
- Thalmayer, A. G., Saucier, G., & Rotzinger, J. S. (2022). Absolutism, relativism, and universalism in personality traits across cultures: The case of the big five. *Journal of Cross-Cultural Psychology*, *53*(7-8), 935-956.
- Thorisdóttir, I.E., Asgeirsdóttir, B.B., Sigurvinsdóttir, R., Allegrante, J.P., & Sigfusdóttir, I.D. (2017). The increase in symptoms of anxiety and depressed mood among Icelandic adolescents: time trend between 2006 and 2016. *The European Journal of Public Health*, *27*(5), 856-861.
- Thórisdóttir, H., & Jost, J.T. (2011). Motivated closed-mindedness mediates the effect of threat on political conservatism. *Political Psychology*, *32*(5), 785-811.
- Topolinski, S., & Boecker, L. (2016). Minimal conditions of motor inductions of approach-avoidance states: The case of oral movements. *Journal of Experimental Psychology: General*, *145*(12), 1589-1603.
- Topolinski, S., Maschmann, I. T., Pecher, D., & Winkielman, P. (2014). Oral approach-avoidance: Affective consequences of muscular articulation dynamics. *Journal of Personality and Social Psychology*, *106*, 885-896.
- Vishkin, A. (2022). Queen's gambit declined: The gender-equality paradox in chess participation across 160 countries. *Psychological Science*, *33*(2), 276-284.
- Vishkin, A., Slepian, M.L., & Galinsky, A.D. (2022). The gender-equality paradox and optimal distinctiveness: More gender-equal societies have more gendered names. *Social Psychological and Personality Science*, *13*(2), 490-499.
- Williams, J. E., & Best, D. L. (1990). *Measuring sex stereotypes: A multinational study* (Rev. ed.). Sage Publications, Inc.
- Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. In *Advances in experimental social psychology* (Vol. 46, pp. 55-123). Academic Press.

- Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3), 363-373.
- Yarkoni, T. (2022) The generalizability crisis. *Behavioral and Brain Sciences*, 45, e1: 1-78.
- Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4), 1036-1040.

Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Social Sciences 209*

Editor: The Dean of the Faculty of Social Sciences

A doctoral dissertation from the Faculty of Social Sciences, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-496532



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2023