Doctoral thesis submitted to
the Faculty of Behavioural and Cultural Studies
Heidelberg University
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy (Dr. phil.)
in Psychology

Title of the publication-based thesis

Information Sampling in Impression Formation Tasks

presented by
Johannes Prager

year of submission
2021

Dean: Prof. Dr. Guido Sprenger
Advisor: Prof. Dr. Klaus Fiedler

Abstract

All three papers included in this thesis, rely on an exemplar trait-impression formation paradigm (Asch, 1946), but throughout each paper different theoretically important aspects of the sampling task are analyzed. Prager, Krueger and Fiedler (2018) elaborate a possible solution to the "less-is-more" debate, that is the question whether more information leads to more or less extreme judgments. The correlation between sample size and how amplified judgments are depends crucially on how the respective sample was stopped. When sample size was experimenter-determined (random), judgments tended to be more conservative for small than for large samples. In contrast, judgments on small samples were polarized for self-truncated sampling (i.e. when the judging participants can themselves decide on when to stop the sampling sequence). Prager and Fiedler (2021b) transferred the self-truncation principle to an inter-group context, where impression targets were groups rather than individuals. By assessing perceived within-group homogeneity in addition to the likeability judgment, we could demonstrate that perceived homogeneity is part of the self-truncation principle: Early truncated samples are not only more polarized but also more homogeneous than samples that are expanded further. Given that out-groups are associated with smaller information samples, these self-truncation effects might constitute a sufficient explanation of out-group homogeneity and out-group polarization.

In Prager et al. (2018) and Prager and Fiedler (2021a) we apply different versions of a yoked controls design: Whereas a primary participant engages in self-truncated sampling, a secondary yoked control receives exactly the same samples passively. This procedure results in regression: Small samples are perceived less polarized for the yoked control than for the primary self-truncating participant. This difference is an exclusive result of different cognitive processing of the sampling input, since the yoked controls design keeps the sampling input itself identical.

All three papers analyze the impact of diagnosticity: Especially negative and extreme information is more diagnostic than positive and moderate input. Highly diagnostic input results in earlier truncation and more polarized judgment. The diagnosticity concept is further elaborated by considering multi-dimensional density in Prager and Fiedler (2021b).

Acknowledgements

This thesis would not have been possible without the support of these people: Klaus Fiedler not only supervised this thesis, he also taught me how do do proper psychological research, how to write papers, how to present ideas and how to survive in the academic world. Thank you for many tough but very insightful discussions and for being an infinite source of creative research ideas. With Linda McCaughey, I could discuss research issues in detail. Thank you for your suggestions, for helping me to sort and position thoughts and ideas in the research projects and for many hours of proof-reading. I got extremely valuable feedback and advice in several stages of the project from Gaël Le Mens, Florian Kutzner, Joachim Krueger, Chris Harris, Thomas Woiczik, Christian Unkelbach and Hans Alves (and everyone else I forgot to mention here). This involved feedback on paper-drafts, comments and discussions in conference presentations and lab meetings. Nicola Ziegler took a critical perspective from outside the psychology bubble. Thank you for questioning the right issues, your sincere interest and your patience. Finally, I want to especially thank Birgit Ehrmann-Prager for supporting me throughout the entire educational process from the very beginning.

List of Publications Included in this Thesis

The following research articles are part of this publication-based thesis:

(A) Prager, J., Krueger, J. I. & Fiedler, K. (2018). Towards a deeper understanding of impression formation – new insights gained from a cognitive-ecological perspective. *Journal of Personality and Social Psychology, 115*(3), 379—397.

(B) Prager, J. & Fiedler, K. (2021). Forming impressions from self-truncated samples of traits – interplay of Thurstonian and Brunswikian sampling effects. *Journal of Personality and Social Psychology. 121*(3), 474–497.

The third article is unpublished, but was submitted to the Journal or Personality and Social Psychology:

(C) Prager, J. & Fiedler, K. (2021). *Small-group homogeneity: A crucial ingredient to inter-group sampling and impression formation.*

The third chapter of this thesis ("Thurstonian Sampling") is a shortened and edited version of the book chapter "Thurstonian Uncertainty in Self-Determined Judgment and Decision Making" (Prager et al., in press). All other chapters are original work.

Contents

Judgments and Decisions from Information Samples

Information samples have traditionally been the basis of judgment and decision tasks. Such tasks involved for example estimating the mean or variance of a deck of numbered cards or to infer from which of two bookbags the experimenter draws poker chips of different color (see Peterson & Beach, 1967 for a summary). More recent approaches to information sampling (see Fiedler, Juslin & Denrell, in press for an overview), focus on how sampling task features, stochastic principles and their interaction with task goals constraint the possible outcomes of cognitive processing and consequent judgments and decisions.

Almost any judgments and decisions refer to latent attributes or dimensions that are in principle not directly and exhaustively accessible by experience. Such latent attributes could be the precise probability of a lottery outcome, student ability, true diagnosis, or likeability. Information samples, experience, but even summary statistics, can merely approximate the true parameters. Human agents (but also computer algorithms and even census data) can never fully access the entirety of relevant instances of proximal observations of these latent dimensions or states (like lottery outcomes, grades in examinations, symptoms, social interaction). Information samples do not contain an imperative that directly results in an action. Thus, the judging agent must integrate potentially diverging and fluctuating information and translate these aggregates into decisions like stopping information search, switching to another information source, choosing an option or to report a final judgment. Such inferences from samples inevitably involve a two-stage process: Samples drawn from the ecology need to be processed, aggregated, and transformed into actions by the cognitive system.

A Cognitive-Ecological Perspective on Sampling

The cognitive-ecological perspective on inference from sampling (Fiedler, 2000; Fiedler & Wänke, 2009; Kutzner & Fiedler, 2017) considers that latent features and characteristics of the environment need to be inferred from proximal stimulus samples. These stimulus samples are the basis of any cognitive processing, judgments and decisions – the sampling ecology must be considered antecedent to cognitive processes. This seemingly trivial notion has crucial consequences on how to analyze judgment and decision behavior: Before we turn to cognitive causes in explaining and predicting judgment and decision patterns, like encoding and retrieval, cognitive strategies, aggregation rules or personal motives, we must consider characteristics of the sampling ecology. Biased judgments (e.g. on out-groups or minorities) are hard to ascribe to personal motives or biased cognitive processing, when already the ecological informational basis of these judgments is skewed, systematically selected or censored. To be clear: Explanations of behavioral phenomena by environmental sampling properties rather than biased cognition do not at all exclude the possibility of subsequent cognitive biases in processing and storing the sampled information or the impact of social or personal motives. Yet, current research taking a cognitive-ecological perspective has demonstrated in various contexts, tasks and paradigms, that many social-cognitive phenomena can be sufficiently explained without assuming any systematic biases in cognitive processing. For example, recent sampling approaches have provided sufficient explanations to phenomena like illusory correlations (Denrell & Le Mens, 2008; Fiedler, 2000), confirmation bias (Fiedler et al., 1999), overconfidence (Erev et al., 1994), loss aversion (Walasek & Stewart, 2015), or social influence (Denrell & Le Mens, 2007).

The thesis-related research articles also build on the sampling approach while considering sample size effects (starting form the less-is-more debate), valence asymmetries and out-group homogeneity. However, neither these, nor many other publications of the sampling framework stop at a mere reproduction of known empirical phenomena by sampling rather than blaming cognitive biases. The sampling perspective provides a deeper understanding of respective judgment and decision phenomena. Knowing that sample properties determine subsequent judgments and decisions, experimentally controlling the constraints of the sampling environment must lead to predictable changes in related judgments and decisions. Controlling sampling constraints and considering the consequences on resulting samples goes far beyond stochastic exercises: Researchers have successfully controlled phenomena in various contexts, where other approaches (especially cognitive biases) have a hard time explaining the resulting specific boundary conditions. Walasek and Stewart (2015) for example demonstrated that loss aversion is strongly impacted by the range-frequency property of gains and losses in the ecology (in usual environments gains are large and infrequent, losses are frequent and small). Exposing participants to a counter-natural environment makes loss-aversion disappear. In a similar approach, Alves et al. (2015) demonstrated that reversing the positive-negative asymmetry in density of words, removes the recognition-advantage of negative words over positives. Or, more recently, Harris et al. (2020) showed that initially biased views on two options either persists or are corrected, depending on whether outcomes in this task are predominantly gains or losses.

Sampling Task Constraints

These examples demonstrate that a comprehensive insight to judgment and decision phenomena is only possible when considering, understanding and controlling the sampling task constraints, that is the way in which the ecology translates into an information sample and the way this sample reaches the human judge. The research included in this thesis will first focus on the less-is-more debate as the questions of how sample size and impressions correlate, but also on the positive-negative valence asymmetry in impression formation and on the out-group homogeneity effect.

But before turning to these phenomena and how they can be tested in experiments, we should acknowledge that research on sampling-based judgment and decision making is a very diverse field: Researchers have relied on a rather large variety of task features. Experimental paradigms can be meaningfully classified in the cognitive-ecological framework by considering the sampling features and constraints. Participants of a sampling task might for example be asked to either chose between options, or to estimate an aggregate value inferred from the sample. Or, they might either observe the sample in order to gain maximum knowledge on the target, or to actually consume the sample, that is facing the consequences of sampling input (like gain and loss or pleasantness and unpleasantness). Similarly, the whole sample might be determined externally, for example drawn by the experimental program, or participants might have influence of some kind on what they see. They might decide on when to stop an unfolding sample or to decide for themselves whether they want to switch sampling from one to another available option.

There are many more features by which sampling tasks and constraints can be classified. But rather than providing an exhaustive taxonomy of sampling tasks, I want to demonstrate that sampling task constraints are much more than adding a label to an experimental paradigm. Sampling task constraints determine the way participants can draw inferences from the sample. Thus, specific technical task features determine unique and theoretically meaningful features of the problem space. Consequently, a different sampling ecology calls for different kinds of cognitive inference and inference-strategies and most likely results in distinct behavior and responses. When we for example consider whether the task is estimation of a target feature from a sample, or whether the task is to chose between two or more choice options, identical sampled input gets a different meaning. Choice tasks require discriminating between options: information that maximizes discriminability (i.e. minimum within-option variance, maximum between-option variance) is most helpful. In contrast, in feature-estimation tasks (like impression-formation tasks), information is helpful and influential when it helps to locate the target on the feature-scale with high confidence, that is when samples appear stable and non-ambivalent about the placement of the target on the feature scale.

Most theoretical work in the three papers involved in this thesis originates from the task feature of self-truncated sampling. The small technical detail of whether participants can decide for themselves when to truncate a sequentially unfolding sample, rather than having to rely on a computer-determined sample, causes manifold consequences on the cognitive-processing level, which even reflect back on the sampling ecology. But, let us consider the details of the impression-formation task, before returning to these cognitive-ecological consequences.

Sampling Task Constraints of the Basic Impression Formation Paradigm of this Thesis

All three research articles included in this thesis rely on variants of a basic person impression formation paradigm. Inspired by Asch (1946), we presented participants with a sequence of traits characterizing a target person. After traits were presented, participants were asked to report their likeability on the target person (Prager & Fiedler, 2021a; Prager et al., 2018). For the experiments in Prager and Fiedler (2021b) the sampling procedure remained identical, but targets were groups rather than individuals and impression ratings aimed at perceived within-group homogeneity in addition to likeability. This experimental paradigm can be classified as impression formation on one target at a time (rather than choice between options).

This specification of task and sampling constraints is much more than a mere taxonomy and labelling empirical effects. Prager et al. (2018) demonstrated that simply changing the sampling constraint of how and when a sequentially unfolding sample is stopped, can reverse sample size effects. By turning sample size from an independent variable in externally (random) determination of sample size into a dependent variable when participants can decide for themselves when to stop the sequence of traits, we changed the inference problem. These theoretically predictable and empirically robust consequences of changing sampling constraints allowed for a new interpretation of the so-called "less-is-more" debate (see next chapter).

Besides establishing a small-sample polarization effect in impression judgments from self-truncated sampling, the aspect of self-truncation makes this research part of a new branch of cognitive-ecological sampling. Dynamic iterations rather than a linear two-stage process must be considered when predicting the sampling and judgment outcomes of self-truncated sampling. For each piece of information being sampled, the judging agent needs to integrate the new information and subsequently decide whether to continue sampling or to stop and to move on to the final judgment. This iterative loop of sampling from the ecology – integration of new information – decision to continue or stop must be repeated until the decision to stop is made. Crucially, this iterative process is recursive: How novel information is processed and whether the sample is continued or truncated depends on the entire sequence of sampling – processing iterations on the current target.

Two further sampling phenomena follow from this dynamic cognitive-ecological perspective. First, we must consider that sampling decisions and impression judgments are not only determined by the mere stimulus input, but also by oscillation (or noise) within the mind of the processing individual. Truncation decisions made by one participant cannot be expected evoke the same level of clarity and preparedness in a yoked-control participant who passively receives the same sample, including the other's truncation decision. Such a passing on of samples between yoked control participants causes regression in the amplification of small-sample polarization effects (Prager & Fiedler, 2021a; Prager et al., 2018, see the chapter on Thurstonian sampling). Second, the self-truncation context would not only imply small-sample polarization, but also small-sample homogeneity. Thus, transferring the small-sample consideration to an inter-group context, Prager and Fiedler (2021b) demonstrated that this is sufficient in producing typical inter-group phenomena, namely out-group homogeneity and out-group polarization (or derogation when considering the valence asymmetry caused by diagnosticity; see chapter on diagnosticity).

All described theoretical considerations and empirical phenomena closely relate to the manifold consequences of the task-feature of self-truncated sampling. It seems worthwhile to elaborate on self-truncation in detail before I discuss the consequences resulting from the presence of Thurstonian uncertainty in self-truncated sampling.

Self-Truncated Sampling in Impression Formation

Impression formation is truly ubiquitous in social environments and social interaction. Typically, it has been assumed that impressions are antecedent to choice – impressions inferred from the observed sample express an approach-avoidance tendency, which can be translated into a specific decision. But even in the absence of a consequential decision or choice, impression formation can be meaningful in itself (e.g. grading students, ranking job applicants or gaining a first impression of a newly encountered person).

Coming back to the trait-sampling paradigm, the distinction of impression formation from choice tasks becomes highly relevant when considering the decision of when to stop a sequentially unfolding sample in self-truncated sampling. We assume that samples are truncated whenever the current impression is clear-cut and sufficiently conflict-free and that ambivalence and contradiction within the sample cause continuation of the sampling sequence. Or, in other words, sampling is truncated when placement of the target person on the likeability scale is easy and consistent. This way of truncating a sample is different from a choice task, where high contrast between choice options leads to sample truncation. Also the epistemic nature of the task (as people do not actually consume the sampled content) is necessary to assume such a truncation-contingency: Hedonic environments would evoke truncation whenever participants face an unpleasant sample and inhibit truncation for pleasant encounters (Denrell, 2005; Fazio et al., 2004).

The assumed truncation-contingency in self-truncated sampling has a crucial impact on the role of sample size: When sample size is fixed and determined by external factors, there is no causal relation between sample content and sample size – sample size remains an independent variable. In self-truncated sampling however, sample size becomes dependent on participant's truncation decisions. As lined out above, we assume a contingency between sample truncation and sampled content. Exactly this central element of how sample size is determined, becomes the key to gaining a new perspective and a possible solution to the less-is-more debate.

A New Perspective on the Less-Is-More Debate

The "less-is-more" principle in interpersonal impression formation is strongly propagated by Norton et al. (2007, 2013). In some of their studies, they relied on an impression formation task, where target persons were characterized by trait words like "ambitious", "bright", "polite", "stubborn", or else. Norton et al. (2007) found a negative correlation between sample size (i.e. the number of traits used to characterize the targets) and impressions. That is, targets were liked less when described by more traits. Ullrich et al. (2013) however argued that such an assumption would contradict statistical principles, like averaging (Anderson, 1965). Given that sampling is stochastically independent, the sample mean is an unbiased estimator of the true population mean. Although we must expect larger fluctuation for small rather than large samples, the mean expected value of samples of any size remains unchanged. Relying on the same impression-formation paradigm and on trait samples as stimulus materials, Ullrich et al. (2013) found no evidence for any kind of sample-size effect.[1]

Solving the "less-is-more" debate is not only of empirical or phenomenological interest. And "less-is-more" does not only violate statistical principles (Ullrich et al., 2013). Many social-cognitive (Zajonc, 1968), but also sampling explanations rely on a "more-is-more" assumption. From the very beginning, judgment and decision researchers assumed that human judges follow (at least approximately) a Bayesian impression-integration strategy (Edwards, 1965; Peterson & Beach, 1967). Given that judging agents do not have any precise expectation on the target before observing information (De Finetti, 1937), larger samples carry stronger evidence and are thus expected to have a greater potential of resulting in a polarized impression, whereas small samples are determined by prior belief, which causes moderate and cautious judgments when priors

[1] Ullrich et al. (2013) focused on the Asch (1946) impression formation task. Norton et al. (2007) also used a romantic dating scenario, which can be even easier understood from a cognitive-ecological perspective. A decision for going on a date with someone is an instance of a very positive impression. Such an impression, that is also based on a small sample will almost inevitably be corrected by sampling more. Selection of extreme events (such as dates or crises) will inevitably be subject to profane regression.

are uninformed (or at least approximately central and symmetrical). In this tradition of considering stochastic properties of samples, Fiedler (2000) assumes that the stronger evidence carried by larger compared to smaller samples causes stronger impressions, which results in illusory correlations (Fiedler et al., 2002, for further examples).

When we strictly follow Norton et al. (2007) and, for a moment, naively interpret the "less-is-more" phenomenon as a universal principle, we end up with a paradoxical situation. Assuming that impressions are antecedent to decisions, by providing a position on an approach-avoidance scale, how can people develop longer lasting relationships? Almost every relationship would constantly deteriorate. The key to solving (at least part of) the contradiction of this conflict is taking sampling constraints into account. In general, statistical rules such as the law of large numbers (Bernoulli, 1713) or Bayesian updating from indifference (De Finetti, 1937) clearly support a general "more-is-more" expectation on the relation between impressions and sample size: Generally, more information renders impressions more confident and sometimes more polarized. This general "more-is-more" principle however has certain boundary conditions. Fiedler and Kareev (2006) for example demonstrated that given a satisficing threshold-based strategy, small samples can (for specific parameter settings) improve contingency detection.

It is crucial to solving the "less-is-more" problem, to consider whether samples are randomly determined and sample size represents an independent variable. Or, whether judging agents can themselves determine when an unfolding sample stops, turning sample size into a dependent variable. For the given person-impression formation from traits paradigm, we allow some participants to truncate the sample themselves whenever they feel to have seen enough in order to give a judgment on the target (Prager et al., 2018). Others must accept the experimenter-truncated sample size (which is chosen randomly). In order to understand sample-size effects in a fixed-sample-size environment and a self-truncated-sample-size environment, we need to combine statistical sampling properties with the impacts of applying a truncation rule. We must expect different sample properties from small and large samples. In addition, in self-truncated sampling, the truncation rule (i.e. the instance of when a sample is stopped) needs to reflect a moment when the sample conveys a clear-cut and conflict-free impression.

Concerning the stochastic properties of small and large samples, the law of large numbers (Bernoulli, 1713) states that for very large samples, the sample parameters approximate the latent population parameters. That also implies that small samples in turn have a greater potential to vary and fluctuate. Small samples more likely amplify an existing trend compared to large samples (Hertwig & Pleskac, 2010). Also, small samples are more likely to convey an unrealistically clear-cut and convergent impression, just by mere sampling error. Applying the proposed sampling rule, it is exactly those instances of clear-cut and conflict-free samples, that are truncated at an early stage. In contrast, samples that are ambivalent and contain conflicting evidence are continued. However, instances of extremely or even exaggeratedly low conflict are much less likely for expanded samples. This contingency of sample size on the clarity of evidence typically leads to a small-sample polarization effect: Impressions on small samples are stronger (one might say more extreme) than impressions on larger samples.

Taking a cognitive-ecological perspective on the issue, we came up with theoretically deduced task constraints, which would lead to either stronger evidence with larger samples (when sample size is fixed externally) or to stronger judgments for small than large samples (when samples are self-truncated). Considering the sampling environment and constraints thereof, "less is more" vs. "more is more" is not an actual contradiction any longer – we can rather specify environmental features that determine the direction of sample size effects on impression judgments. However, to respond to the question, whether "less" is really "more", we need to consider further constraints, namely the cost of errors versus the benefit of sampling less. I will discuss these aspects as an outlook in the last chapter – in the three research articles included in this thesis, we primarily focused on the peculiarities of the sample-size-impression-strength dependency.

The sample truncation rule is not peculiar to human judges, but might as well be expressed in statistical terms. In Prager and Fiedler (2021a) and Prager and Fiedler (2021b) we give examples of simplified statistical optional stopping rules. Truncation of the sample at a moment of clarity and perceived stability might be expressed as a sufficiently small standard error as in Haldane's (1945) labour saving method of sequential estimation. Or, when using Bayesian methodology, truncation is contingent on the width of a fixed-weight posterior highest-density-interval. Such an interval covers the narrowest possible range of, say 90%, posterior probability. Both exemplar statistical optional-stopping procedures result in a negative relation between sample size and the strength (i.e. polarization) of resulting calculated estimates (which are meant to imitate human impression judgments) for the vast majority of possible parameter settings.

Self-Truncated Sampling: A Sufficient Condition to Out-Group Homogeneity

In Prager and Fiedler (2021b) we argued that it is exactly those consequences of self-truncation that form a sufficient condition to produce out-group homogeneity and out-group polarization effects. For elaborating on this claim, it is necessary to transfer the self-truncation principles to the inter-group context. In Prager and Fiedler (2021b) we consider groups rather than individuals as impression targets, while the self-truncated sampling from traits procedure remains the same. Each sampled trait is assigned to a different member of the current target group. Besides likeability, which aims at the central moment of the target impression, we assessed perceived homogeneity, addressing how likeability is dispersed within the target group. We confirmed that, given the epistemic impression formation sampling goal, self-truncated trait sampling results in the already described small-sample polarization effect as well as in a small sample homogeneity effect. Assuming that it is typically the out-groups that we know and sample little about (Linville et al., 1989; Park & Rothbart, 1982), we can transfer these sample size effects to the inter-group context.

By considering the consequences of self-truncated sampling, it is possible to provide sufficient reasons to solving the "less-is-more" debate, but also to account for out-group homogeneity effects. In order to understand self-truncated sampling however, we must take exclusively cognitive processes in addition to the sampling ecology into account. Self-truncated sampling is a dynamic interactive process of sampling from the ecology and impression-updating within the mind of the judging individual. When considering the moment of sample truncation for example, it is not only the clarity of evidence that triggers truncation, but also a simultaneous amplification of this impression by intra-cognitive processes.

Thurstonian Sampling

Brunswikian and Thurstonian Sources of Uncertainty[2]

The concept of subjective scaling (Thurstone, 1927) has now been used and applied in different contexts for almost a century. Nevertheless, we can still gain meaningful and novel insight into sampling and judgment behavior in impression formation tasks from taking a Thurstonian perspective on sampling tasks. The first two articles included in this thesis (Prager & Fiedler, 2021a; Prager et al., 2018) apply the concept of Thurstonian uncertainty to impression formation from self-truncated sampling. Crucially, the dynamics of self-truncated sampling make Thurstonian uncertainty an integral part of sampling and judgment behavior.

Thurstonian uncertainty refers to the notion of a dispersed and fluctuating mental representation of the judgment target, even when the stimulus input is held perfectly constant (Thurstone, 1927). Each individual impression might vary over time, context and between individual judges. In contrast, Brunswikian uncertainty (for a detailed definition an distinction, see Juslin & Olsson, 1997) covers all aspects of fluctuation, incompleteness or sampling error caused by the mere stimulus materials. Before analyzing the characteristics and consequences of these two distinct determinants of sampling behavior and judgment, it is worthwhile to elaborate more on the conceptualization of Thurstonian uncertainty, starting from Thurstone's (1927) idea of psychological scaling of stimuli.

A Law of Comparative Judgment

By proposing a "law of comparative judgment", Thurstone (1927) provided a method for scaling stimuli on a latent attribute dimension. For the person-impression formation

[2] This chapter is a shortened and adapted version of the chapter "Thurstonian Uncertainty in Self-Determined Judgment and Decision Making" in Fiedler et al. (in press)

paradigm of this thesis' studies, the latent attribute dimension is likeability and the to be scaled stimuli are traits (e.g., "honest", "cruel", "creative"). Thurstone's core assumption is that traits (or all other kinds of stimuli) are not represented as unchanged scalars on a fixed likeability dimension, but rather that each stimulus takes a normally distributed scale position that fluctuates dependent on time and situation. The latent likeability dimension on which different traits take their normally distributed positions can thus only be inferred from paired comparisons over different instances. Thus, once the scale has been set by a pairwise comparison of two traits, the relative scale positions of further stimuli can be determined from preference judgments of these over the original stimuli.

While Thurstone's (1927) model of psychological scaling relied on paired-comparison data, his general notion of a distributive (and therefore fluctuating) representation of target stimuli can be applied to modeling judgments of all kinds. Impression formation from person traits calls for the integration and forming an aggregate evaluation rather than their pairwise comparison. Yet, Thurstone's idea of a relative latent psychological scale is still applicable: Locating target stimuli on a common scale is the first step of inferring an aggregate likeability impression. Most importantly, likeability judgements that were combined from a sample of traits fluctuate from one situation to another.

Brunswikian and Thurstonian Sources of Uncertainty

The ambiguity, fluctuation and uncertainty resulting from Thurstone's concept of stimulus scaling has important implications for judgments and decisions inferred from incomplete and indirect stimulus samples. Based on the law of comparative judgment, Juslin and Olsson (1997) classified "Thurstonian" in contrast to "Brunswikian" uncertainty. Brunswikian sources of uncertainty are fully environmentally determined. As Brunswikian uncertainty is caused by the incompleteness, invalidity and insufficiency of a sample in characterizing the true population from which it is drawn and related true parameters, this kind of uncertainty can in principle not be reduced by the cognitive system through more careful assessment or enhanced processing capacity. Even unbiased, lossless and flawless statistical procedures or perfectly operating algorithms are subject to Brunswikian uncertainty.

Thurstonian sources of uncertainty, in contrast, exclusively take place within the mind of the processing individual. Thus, Thurstonian uncertainty can cause varying judgments between different contexts, individuals and situations, even when Brunswikian uncertainty is kept constant (i.e. when information or the input stimuli remain unchanged). Oscillations related to Thurstonain uncertainty can result for different reasons, such as memory responses or interference, or variability inherent to the perceptual and nervous system.

Although both types of uncertainty in sampling tasks are clearly distinct conceptually and refer to either ecological (Brunswikian) or intra-cognitive (Thurstonian) influences, they are neither mutually independent nor do they reflect additive or separable consecutive stages of an overarching process. Judgment tasks cannot be split into categories, where only Brunswikian or Thurstonian uncertainty effects responses. Rather, the interplay of ecological and cognitive sampling processes reflects an intertwined iterative process. That is especially the case when judging individuals can base their judgments on self-truncated samples, when they can themselves determine the moment of stopping the sequentially unfolding trait sample in the impression-formation task. In such an iterative process of sampling (vs. stopping) and impression updating, Brunswikian uncertainty can produce accentuated and clear-cut evidence, which is especially likely for small samples. Or, also as a consequence of Brunswikian uncertainty (i.e. sampling error), an initial sample might contain conflicting and ambivalent evidence. The observed evidence (trait sample) triggers associative and generative cognitive processes within the mind of the judging individual, extracting the contextual and individual meaning of the current state of information. Thus, Thurstonian uncertainty might as well result in either enhanced clarity and convergence, amplifying the Brunswikian trend, or it might result in a conflicting and contradictory interpretation of the sampled evidence. Thus, in impression formation from self-truncated samples, updating one's impression involves an iterative process of repeated interaction between environment (the Brunswikian trend) and mind (interpretation and integration of what has been observed). Both, the moment of sample truncation (sample size) as well as the final impression judgment are a product of this repeated interaction.

Sample Truncation

As explained in earlier chapters, impression formation calls for a truncation strategy, which aims at placement of the target (person/group) on a target dimension (likeability). It seems sensible to assume truncation to depend on sufficient settlement, stability, and freedom of conflict. Thus, the moment of truncation can be characterized as a situation where the sampled evidence is integrated into a stable impression, that is not expected to change much when new evidence is added. Truncation takes place, when the anticipated informational value of newly added traits is expected to be minimal.

Stochastic indeterminacy and conflict versus settlement and expected stability in the Brunswikian sample can be expressed by statistics of the sampled values (e.g., trait valence values). Those could be the standard deviation, or the width of the posterior highest-density-interval in Bayesian updating. In Prager and Fiedler (2021a) we based our demo-simulation on these statistical approaches to optional stopping of sequential samples (Edwards, 1965; Haldane, 1945).

In our characterization of the sampling process as a genuine interaction of Brunswikian and Thurstonian uncertainty, statistical stopping rules that exclusively consider the Brunswikian sample do indeed provide a first access to self-truncated sampling, but they ignore Thurstonian uncertainty in the mind of the judging individual. Both, the decision to continue sampling or to stop and the final impression are inevitably impacted by Thurstonian uncertainty. The moment of truncation cannot be expected to be exclusively determined by the clarity and convergence of the Brunswikian sample, but also by an alignment with a Thurstonian interpretation of this evidence that supports the perception of clarity and freedom of conflict. In other words, samples are truncated when sample and mind align to a clear-cut impression. Similarly, resulting (likeability) judgments are strongly affected by both Brunswikian and Thurstonian evidence. All traits used as stimulus materials were pre-tested for valence. All three papers included in this thesis show a very strong determination of likeability judgments by the simple average of pre-test valence values of actually sampled traits. Yet, this research simultaneously demonstrates that participants systematically underweight input that has little informative value (e.g., when it is redundant to what is already known) and overweight instances of highly informative input (see the following discussion on *diagnosticity*).

In previous research, the interplay of Brunswikian and Thurstonian uncertainty in self-truncated sampling has mainly been examined in choice tasks (Busemeyer & Townsend, 1993; Ratcliff, 1978), but not in impression formation contexts. The assessment of Brunswikian and Thurstonian sampling by Prager et al. (2018) and Prager and Fiedler (2021a) thus goes beyond a mere technical refinement of already existing experimental paradigms. Rather, the small changes in technical details lead to considerable differentiation in what is to be expected from sampling and judgment behavior theoretically. The full potential of theoretical contribution of adding the concept of Thurstonian uncertainty to obvious Brunswikian uncertainty of the sampled content, unfolds when considering the yoked controls design.

Detecting Thurstonian Sampling: The Yoked Controls Design

Since Thurstonian uncertainty covers a multitude of cognitive processes (perceptual, memory-related, neuronal and all variants thereof) it seems reasonable to conceptualize Thurstonian uncertainty as normally distributed random noise scattered around the Brunswikian center point. This highly simplified conceptualization of the outcomes of a multi-causal process (Galton, 1894) comes close to the dispersion of stimuli on the latent cognitive scale in Thurstone's (1927) original model.

Relying on this simplified "black box" perspective on Thurstonian uncertainty, we applied the *yoked controls* design in a series of impression formation experiments (Prager & Fiedler, 2021a; Prager et al., 2018). A primary judge in a pair of participants could sample traits and truncate the sample when they felt ready to form a judgment. The secondary participant received exactly the same trait sample, presented in the same order and limited by the primary judge's truncation decision, making them the yoked control. This arrangement is especially suited for detecting the impact of Thurstonian uncertainty, since Brunswikian uncertainty is kept identical between the yoked participants: they receive exactly the same traits in the same order and up to the same sample size. The states of mind and cognitive processing (i.e. Thurstonian uncertainty) however cannot be expected to be equally synchronized. The moment of sample truncation was only tuned to the primary participant's mind, not to the presumably different Thurstonian evaluation of the second participant. The primary participant most likely experienced synchrony between Brunswikian and Thurstonian processes, whereas there is more potential for contradiction for the secondary yoked partner, whose mind set is not fully synchronized with the first participant.

The sample size effects caused by self-truncated sampling (see previous chapter) carry over to the secondary yoked partners, who passively observe a formerly self-truncated trait sample. As the Brunswikian sample is kept identical, and since sample truncation effects are regularly also visible in the mere Brunswikian sample content, secondary yoked controls can be expected to show the same tendencies in their dependence of impressions from sample size. However, the misalignment and conflict between Thurstonian uncertainty causes regression in the yoked partners' sample size effects. Yoked control's impressions on small samples are generally weaker than the original self-truncated impressions. The regressive shrinkage was confirmed empirically in Prager et al. (2018) and Prager and Fiedler (2021a).

Critical Tests of Yoked Controls' Regression

In a more sophisticated task variant (Prager & Fiedler, 2021a), we were able to generate experimental conditions where yoked controls showed different levels of dependency between their Thurstonian mind states, resulting in specific levels of regression in their sample size effects. So far, the yoked controls design consisted of yoked pairs of which the primary partner worked on the self-truncated impression formation task, whereas the secondary partner passively received that same sample passively. Although we confirmed the expected regression phenomenon (i.e. the correlation between sample size and the strength/polarization of impression judgments shrank noticeably between the primary to the secondary participants) in the first yoked controls experiment (Prager et al., 2018), there is still one critical test missing. Yoked participants did not only differ in how their samples were truncated (which we obviously assume to be the cause of the regression effect). They also engaged in either active sampling (self-truncated) or passive reception of the sample (yoked control). Our yoked-controls design (Prager & Fiedler, 2021a) presented participants in an other-yoked condition with trait samples that other participants had truncated in a previous block of trials, identical to the former experiments by Prager et al. (2018). In contrast, participants in a self-yoked condition, in the second block received copies of their own trait samples, which they had themselves truncated in the first block.

This extension of the paradigm is a critical test of whether the regression phenomenon is an artifact of active versus passive sampling. Since all participants of the second block engage in passive sampling, we should not observe any difference in regression here when the effect was caused by a task-feature artifact. But the crossed-over design of self- and other yoked controls can provide further theoretical insight than testing against an artifact: The self- and other-yoking conditions decomposed the Thurstonian process into (a) the impact of mere inter-temporal oscillations in mental representations of the same trait samples (self-yoking) versus (b) the joint impact of both inter-temporal and inter-personal oscillations (other-yoking). Thus, Thurstonian uncertainty in the yoked-controls trials is highly dependent on Thurstonian uncertainty of the previous self-truncated sampling trials for participants who received their own samples again. In contrast, other-yoked participants' Thurstonian uncertainty is much less dependent due to changing the judging agent between blocks. The empirical results confirmed this expectation: Regressive shrinkage was indeed stronger in the other-yoked than in the self-yoked condition. Asking the same person to form impressions from the same trait sample twice, separated only by the delay between blocks, causes less asynchrony in Thurstonian preparedness to judge than asking different persons.

Diagnosticity: A Systematic Interaction of Brunswikian and Thurstonian Properties

All three thesis papers (Prager & Fiedler, 2021a, 2021b; Prager et al., 2018) confirm that aside from the measurement of unspecific Thurstonian oscillation by the yoked controls design, sampling behavior and impression judgments are driven by a systematic and predictable interaction of Brunswikian and Thurstonian properties. The Brunswikian stimulus input is to be processed and aggregated on a meaningful background of task and contextual features. Even when Brunswikian uncertainty remains unchanged, identical stimuli can change their meaning and thus their informativeness and *diagnosticity* considerably when they are processed or integrated under different sampling or task goals or in different contexts. In the following section, I will classify and characterize the most important features of traits in the context of a person impression formation task, namely positive versus negative and moderate versus extreme valence in addition to density. The stimulus features are obviously part of the Brunswikian sample. Task and context features are not actually dependent on the Brunswikian sample, but rather represent systematic Thurstonian impacts. Consequently, exchanging those features, by for example switching the sampling goal, must be expected to change the Brunswikian information's meaning and thus the weight given to respective pieces of information. However, before discussing the possibility of dynamic context effects, it is necessary to consider diagnosticity in the present impression formation context first.

Diagnosticity

Diagnosticity has been considered by two streams of theorizing and research: First, diagnosticity is an important property of information integration in a Bayesian perspective on judgment and decision making. Secondly, diagnosticity has regularly been considered by social-cognition researchers. I will try to unify these perspectives here by demonstrating that social-cognitive diagnosticity effects can be explained and predicted in the conceptual framework of the original Bayesian updating framework.

In classical decision theory, diagnosticity signifies the change in preference (or belief) caused by newly integrated information (Edwards, 1965). When we consider a classical bookbag problem (Phillips & Edwards, 1966), the task is usually to infer the origin of a sample from one of two bookbags: One bag contains 70% red and 30% blue poker chips, whereas the other bag the reversed proportion of 70% blue, 30% red. The experimenter draws (hidden to the participant) one chip after another from one of the bags and the participant is asked to guess which bag the chips are taken from. In such a context, a sequence of red-blue-red-blue would be very low in diagnosticity. The sequence does not shift the belief into one or the other direction, the likelihood ratio of the data given one over the other hypothesis (70:30 vs. 30:70 bag) is undecisive (i.e. close to 1). In contrast, a sequence of red-blue-red-red would shift the preference towards the 70% red bag, the likelihood ratio differs from 1 (as the data is much more likely given the 70% red hypothesis) – the sequence is diagnostic.

Social-cognition literature was rather considered with valence than poker-chip-proportions in bookbags. Here, negative and extreme behaviors or traits are regularly more diagnostic. It takes, for example, less behavioral observations to confirm a negative impression, but much more observations to disconfirm a negative impression than a positive impression (Gidron et al., 1993; Rothbart & Park, 1986). Generally, negative and extreme information has a greater impact on the resulting impression (Skowronski

& Carlston, 1987). Although the social-cognition and the classical judgment and decision making literature differ in many respects, the view on diagnosticity can be easily unified. We can also describe the diagnosticity-valence effects in terms of Bayesian updating. Negative and extreme observations have a greater potential of changing the impression judgment. For example, both a honest and a dishonest person tell the truth on most occasions. It is only the rare instances of observing someone telling a lie that differentiate honest from dishonest people. The likelihood ratio of observing telling the truth given positive (honest) over negative (dishonest) is rather indifferent, whereas the ratio has a clear tendency towards negative (dishonest) after observing lies.

In all three papers, we empirically tested for the differential diagnosticity of valence (Prager & Fiedler, 2021a, 2021b; Prager et al., 2018). In all experiments, we rely on pre-defined population sets of traits from which samples are drawn. These population sets are constructed to reflect certain valence properties. Sets reflected negative versus positive and orthogonally extreme versus moderate sets. The diagnosticity over the valence range is illustrated in Figure 1. All three articles report convergent evidence on the diagnostic value of negative and extreme compared to positive and moderate valence. This higher diagnosticity manifests in both sampling behavior and judgments. Samples from population sets of high diagnosticity are truncated earlier, they result in stronger impression judgments and are perceived to be more homogeneous.

Density

In Prager and Fiedler (2021b) we relied on the density hypothesis (Unkelbach et al., 2008) to explain the diagnosticity patterns. Informational input (here: trait words) can be scaled on multiple dimensions rather than only one (likeability) target scale. Such multi-dimensional abstract scaling of stimuli allows to determine the distance (or density as opposite distance) between stimuli. Since negative and extreme words are more distant from all other words in the total set compared to positive and moderate words, we can explain diagnosticity effects by density: The uniqueness of negative and extreme traits causes fast and strong formation of an impression judgment.

When we discuss density (or distance as opposite density) in an impression-formation paradigm, we must consider two facets when predicting sample truncation and impression judgments. There is distance within a sample and the distance of a sample as a

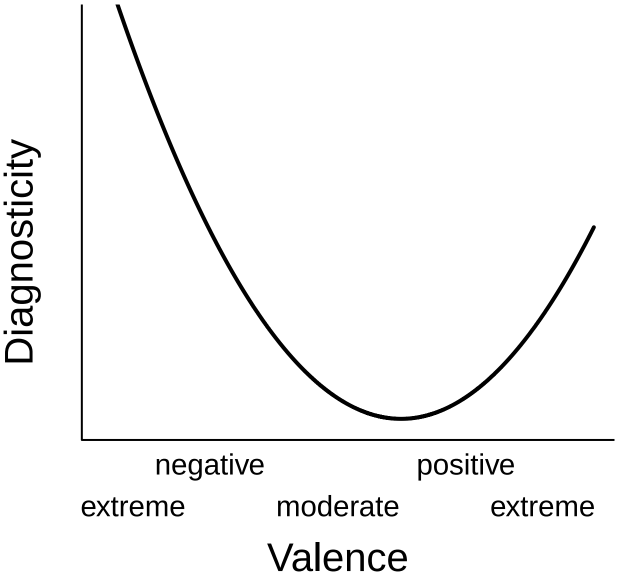


Figure 1: Prototypical diagnosticity for different levels of valence. Negative and extreme valence is expected to be higher in diagnosticity than positive and moderate valence.

whole to the context of other targets. In other words, we must consider *within sample* and *between sample* distance. The two facets have differential impact on sample truncation and judgments: We can expect early truncation, strong judgments and high perceived target homogeneity when within sample distance is low (i.e. high within sample density) and when, simultaneously, between sample distance is high. This prediction becomes clear, when we conceptualize density as redundancy (Soll, 1999). The more similar the traits on different dimensions, the more redundant they are (i.e. features of one trait can easily be inferred from the other trait), the higher the distance, the more unique and individual characteristics they carry. Redundancy is a fundamentally different signal to sample truncation between and within a sample. *Between* sample-redundancy (typical for the positive and moderate domain) is non-diagnosticity. When a sample is (partly) redundant to the context of other samples, for example when a group does hardly differ from other groups we know, the impression hardly changes, the informational value is low, and the situation calls for more information. However, when *within* sample redundancy is high, it is a strong signal of convergence, freedom of conflict and validity. The different pieces of information within the sample confirm each other. That calls for truncation and the judging individual is enabled to form strong and confident impressions (given that the between sample distance allows for strong impressions). Additionally, within sample redundancy should drive the within target homogeneity perception.

In an earlier chapter, I discussed self-truncated sampling as a sufficient cause to out-group homogeneity effects. Now, having considered the crucial role of density, the prediction becomes even more clearly. Density amplifies such an inter-group effect: When group-related samples are of high within and low between sample density, small-group polarization and small-group homogeneity become even stronger and more predictable as we consider a multi-dimensional feature-space rather than the singular dimension of likeability. Furthermore, the typical ecology is skewed in density (as we confirmed for the used trait set in Prager & Fiedler, 2021b): Negative and extreme traits are more distant to all other traits, positive and moderate traits are closer to other traits. This skewness predicts a small-group derogation effect. When samples contain negative and extreme information they are typically high in between sample distance, truncated early and judged strongly and homogeneously.

Discussion

In recent years, researchers' interest in judgments and decisions based on experience (i.e. on samples) has considerably increased. Different lines of research have been established. The three articles included in this thesis demonstrate repeatedly, that considering the boundary conditions and constraints of sampling environments is far beyond discussing technical details. Explaining phenomena "by experience" or to characterize effects as phenomena of "sampling" has little value if researchers are not specific on the sampling constraints that are either inherent to the environment, or that are imposed by researchers. We argued in Prager et al. (2018) for example, that the task constraint of either externally fixed or self-truncated sample size can reverse a "more-is-more" into a "less-is-more" pattern.

Engaging in the debate, we refrained from explicitly stating that "less" is actually "more", but rather focused on the small sample polarization effect: Small samples are judged more strongly. Whether that is "more" in a sense that participants profit from the polarized judgments from small samples, can only be determined by considering further sampling constraints. Furthermore, self-truncated sampling has an unsolvable conflict inherent to sample truncation and judgment. When a small sample contains extreme and conflict-free evidence, it might indicate a strong population trend, or it might as well just reflect an instance of clarity caused by sampling error. When we analyze truncation and judgment behavior, we see – from hindsight – that people's small samples actually intermix population trends and sampling error (and Thurstonian error). But when we consider the situation in judging a currently unfolding sample, we do not know about the population properties and are forced to base the judgment on the sampling input alone. Even sophisticated statistical procedures cannot solve this problem of differentiating trends and error in early clear-cut evidence. Therefore, participants do not need to be naive on the origins of samples to truncate clear-cut samples early (Le Mens & Denrell, 2011). Most circumstances however indicate that it is profitable to exploit such instances of early "good luck" (Edwards, 1965). Whether it is beneficial to go for an early trend or to be more cautious and only accept higher levels of confirmation depends on the cost-benefit context. Time pressure might for example require fast trend-detection (Fiedler et al., 2021). Such changes in the cost-benefit context however can hardly effect the described self-truncation principles. Although sample size-effects fade out for larger sample size and would thus result in approximate zero-effect sizes for accuracy-focused truncation strategies, the self-truncation principles should still hold, with the only difference that evidence-thresholds required for truncation are set higher.

Connected to self-truncation effects, the yoked controls design demonstrates that a simple classification and mere labelling of sampling tasks and sampling constraints without any sampling-theoretical background is insufficient. We can label both, externally fixed sampling as well as yoked controls sampling as "passive" sampling, since in both conditions, participants receive samples in the same presentation mode and cannot decide on sample truncation. As already discussed, yoked controls conserve small sample polarization from the formerly self-truncated samples whereas externally fixed sampling lacks such a contingency. Similarly, when only considering the actual sampling input, self-truncated samplers and yoked control participants do not differ, while considering Thurstonian uncertainty allows for a precise analysis of regression of sample-size effects. Applying the concept of Thurstonian uncertainty to the yoked controls design, allowed us to deduce further predictions: The comparison of self- versus other- yoked controls in Prager and Fiedler (2021a) precisely followed the expected differential dependency when considering the same participant in different occasions or when additionally exchanging the participant.

Only the consideration of meaningful and relevant elements of sampling constraints in a theory-driven cognitive-ecological analysis can lead to meaningful, justified and transferable predictions and explanations in specific sampling contexts. Since inference from samples is omnipresent also in everyday judgment and decision making, it is worthwhile to analyze the impact of sampling constraints in the real-world, too. In any sampling scenario, the causes of sample truncation inevitably impact resulting judgments and decisions: It makes a difference whether samples are a priori fixed, self-truncated, or whether information is communicated from an actively sampling sender to a rather passive receiver in a scenario close to the yoked-controls design.

In other sampling contexts, task constraints might change how judgments can be inferred. We mainly considered a communion-focused likeability judgment. Here, negative and extreme observations proved to be most informative and diagnostic. However, when the task goal is more agency-centered (e.g. in an evaluation of performance rather than likeability) we must expect changes in how sampled information is weighted. For agency contexts, positive behavior is regularly more diagnostic than negative behavior (Skowronski & Carlston, 1987). Thus, the diagnosticity considerations presented in the research papers have clearly defined boundary conditions, that are set by the task goal and the stimulus environment. Concerning our density explanation (Prager & Fiedler, 2021b), we must expect clearly different effects when participants get used to a reversed task environment that does not reflect the usual positive-negative asymmetry.

The research presented in this thesis is part of a recent dynamic approach to sample-informed judgment and decision making. New sampling approaches do not only consider that samples form the interface between the environment and cognitive processing, but also that most often, the cognitive system and the sampling process dynamically interact. In self-truncated sampling, the sample content is still drawn at random, but sample size is systematically determined. Previous sampled content impacts whether or not further evidence is sought. And since such an iterative evaluation and integration of a sequentially unfolding sample inevitably requires the involvement of the cognitive system, Thurstonian uncertainty becomes an integral part of the sampling process. Such inter-dependent iterative structures are typical of real-world sampling scenarios. Systematic theoretical (defining the inference problem) and empirical analyses are worthwhile in explaining and predicting sample-informed judgments and decisions.

Bibliography

Alves, H., Unkelbach, C., Burghardt, J., Koch, A. S., Krüger, T. & Becker, V. D. (2015). A density explanation of valence asymmetries in recognition memory. *Memory & Cognition*, *43*(6), 896–909.

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, *70*(4), 394–400.

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, *41*(3), 258–290.

Bernoulli, J. (1713). *Ars conjectandi: Opus posthumum*. Thurnisii.

Busemeyer, J. R. & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(3), 432–459.

De Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, *7*(1), 1–68.

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review*, *112*(4), 951–978.

Denrell, J. & Le Mens, G. (2007). Interdependent sampling and social influence. *Psychological Review*, *114*(2), 398–422.

Denrell, J. & Le Mens, G. (2008). Illusory correlation as the outcome of experience sampling. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*(30).

Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology*, *2*(2), 312–329.

Erev, I., Wallsten, T. S. & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519–527.

Fazio, R. H., Eiser, J. R. & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology*, *87*(3), 293–311.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*(4), 659–676.

Fiedler, K., Juslin, P. & Denrell, J. (in press). *Sampling in judgment and decision making*. Cambridge University Press.

Fiedler, K. & Kareev, Y. (2006). Does decision quality (always) increase with the size of information samples? some vicissitudes in applying the law of large numbers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 883–903.

Fiedler, K., McCaughey, L., Prager, J., Eichberger, J. & Schnell, K. (2021). Speed-accuracy trade-offs in sample-based decisions. *Journal of Experimental Psychology: General*, *150*(6), 1203–1224.

Fiedler, K., Walther, E., Freytag, P. & Plessner, H. (2002). Judgment biases in a simulated classroom–a cognitive-environmental approach. *Organizational Behavior and Human Decision Processes*, *88*(1), 527–561.

Fiedler, K., Walther, E. & Nickel, S. (1999). The auto-verification of social hypotheses: Stereotyping and the power of sample size. *Journal of Personality and Social Psychology*, *77*(1), 5–18.

Fiedler, K. & Wänke, M. (2009). The cognitive-ecological approach to rationality in social psychology. *Social Cognition*, *27*(5), 699–732.

Galton, F. (1894). *Natural inheritance*. Macmillan; Company.

Gidron, D., Koehler, D. J. & Tversky, A. (1993). Implicit quantification of personality traits. *Personality and Social Psychology Bulletin*, *19*(5), 594–604.

Haldane, J. (1945). A labour-saving method of sampling. *Nature*, (155), 49–50.

Harris, C., Fiedler, K., Marien, H. & Custers, R. (2020). Biased preferences through exploitation: How initial biases are consolidated in reward-rich environments. *Journal of Experimental Psychology: General*, *149*(10), 1855–1877.

Hertwig, R. & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, *115*(2), 225–237.

Juslin, P. & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, *104*(2), 344–366.

Kutzner, F. & Fiedler, K. (2017). Stereotypes as pseudocontingencies. *European Review of Social Psychology*, *28*(1), 1–49.

Le Mens, G. & Denrell, J. (2011). Rational learning and information sampling: On the 'naivety' assumption in sampling explanations of judgment biases. *Psychological Review*, *118*(2), 379–392.

Linville, P. W., Fischer, G. W. & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: Empirical evidence and a computer simulation. *Journal of Personality and Social Psychology*, *57*(2), 165–188.

Norton, M. I., Frost, J. H. & Ariely, D. (2007). Less is more: The lure of ambiguity, or why familiarity breeds contempt. *Journal of Personality and Social Psychology*, *92*(1), 97–105.

Norton, M. I., Frost, J. H. & Ariely, D. (2013). Less is often more, but not always: Additional evidence that familiarity breeds contempt and a call for future research. *Journal of Personality and Social Psychology*, *105*(6), 921–923.

Park, B. & Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. *Journal of Personality and Social Psychology*, *42*(6), 1051–1068.

Peterson, C. R. & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*(1), 29–46.

Phillips, L. D. & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*(3), 346–354.

Prager, J. & Fiedler, K. (2021a). Forming impressions from self-truncated samples of traits - interplay of Thurstonian and Brunswikian sampling effects. *Journal of Personality and Social Psychology*, *121*(3), 474–497.

Prager, J. & Fiedler, K. (2021b). *Small-group homogeneity: A crucial ingredient to inter-group sampling and impression formation* [Manuscript submitted for publication].

Prager, J., Fiedler, K. & McCaughey, L. (in press). Thurstonian uncertainty in self-determined judgment and decision making. In K. Fiedler, P. Juslin & J. Denrell (Eds.), *Sampling in judgment and decision making*. Cambridge University Press.

Prager, J., Krueger, J. I. & Fiedler, K. (2018). Towards a deeper understanding of impression formation-new insights gained from a cognitive-ecological perspective. *Journal of Personality and Social Psychology*, *115*(3), 379–397.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Rothbart, M. & Park, B. (1986). On the confirmability and disconfirmability of trait concepts. *Journal of Personality and Social Psychology*, *50*(1), 131–142.

Skowronski, J. J. & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology*, *52*(4), 689–699.

Soll, J. B. (1999). Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psychology*, *38*(2), 317–346.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286.

Ullrich, J., Krueger, J. I., Brod, A. & Groschupf, F. (2013). More is not less: Greater information quantity does not diminish liking. *Journal of Personality and Social Psychology*, *105*(6), 909–920.

Unkelbach, C., Fiedler, K., Bayer, M., Stegmüller, M. & Danner, D. (2008). Why positive information is processed faster: The density hypothesis. *Journal of Personality and Social Psychology*, *95*(1), 36–49.

Walasek, L. & Stewart, N. (2015). How to make loss aversion disappear and reverse: Tests of the decision by sampling origin of loss aversion. *Journal of Experimental Psychology: General*, *144*(1), 7–11.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*(2, Pt.2), 1–27.

Appendix A

Towards a Deeper Understanding of Impression Formation – New Insights Gained From a Cognitive-Ecological Perspective.

Towards a Deeper Understanding of Impression Formation — New Insights Gained from a Cognitive-Ecological Perspective

Johannes Prager[a], Joachim I. Krueger[b] and Klaus Fiedler[a]

[a]Heidelberg University; [b]Brown University

ABSTRACT

Impression formation is a basic module of fundamental research in social cognition, with broad implications for applied research on interpersonal relations, social attitudes, employee selection, and person judgments in legal and political context. Drawing on a pool of 28 predominantly positive traits used in Solomon Asch's (1946) seminal impression studies, two research teams have investigated the impact of the number of person traits sampled randomly from the pool on the evaluative impression of the target person. Whereas Norton, Frost, and Ariely (2007) found a "less-is-more" effect, reflecting less positive impressions with increasing sample size n, Ullrich, Krueger, Brod, and Groschupf (2013) concluded that an n-independent averaging rule can account for the data patterns obtained in both labs. We address this issue by disentangling different influences of n on resulting impressions, namely varying baserates of positive and negative traits, different sampling procedures, and trait diagnosticity. Depending on specific task conditions, which can be derived on theoretical grounds, the strength of resulting impressions (in the direction of the more prevalent valence) (a) increases with increasing n for diagnostic traits, (b) is independent of n for non-diagnostic traits, or (c) decreases with n when self-truncated sampling produces a distinct primacy effect. This refined pattern, which holds for the great majority of individual participants, illustrates the importance of strong theorizing in cumulative science (Fiedler, 2017) built on established empirical laws and logically sound theorizing.

KEYWORDS

less-is-more, Brunswikian sampling, Thurstonian sampling, primacy advantage, diagnosticity

Drawing on the seminal impression-formation paradigm originally developed by Asch (1946), Norton et al. (2007, 2011), and then Ullrich et al. (2013) investigated evaluative impressions as a function of the number of traits sampled from a basic set. Participants were asked to evaluate a target person described by a sample of traits that were randomly drawn from a universe of 28 traits used in Asch's (1946) seminal work. The majority of traits was positive, reflecting the preponderance of positive (i.e., normative) behavior in reality. The central finding reported by Norton et al. (2007, 2011) was called a less-is-more effect: Smaller samples led to more positive impressions. However, two

Johannes Prager, Department of Psychology, Heidelberg University, johannes.prager@psychologie.uni-heidelberg.de

Joachim I. Krueger, Department of Cognitive, Linguistic & Psychological Sciences, Brown University

Klaus Fiedler, Department of Psychology, Heidelberg University

The work underlying the present article was supported by a Grant provided by the Deutsche Forschungsgemeinschaft to Klaus Fiedler (FI 294/23-1).

follow-up studies by Ullrich et al. (2013), one of which was coordinated with the Norton team to be an exact replication, did not replicate the less-is-more effect. Instead, judgments closely resembled the average valence of all sampled traits, regardless of sample size. A synthesis of all experiments led Ullrich et al. (2013) to conclude that an unbiased averaging algorithm – in line with Anderson's (1981) information-integration model – provides a satisfactory account of the entire evidence, including the seemingly divergent findings reported by Norton et al. (2013).

Ullrich et al. (2013) only relied on aggregation of all available data, not on substantial theorizing about underlying mechanisms. Although they discussed several reasons why evaluations might decrease, remain constant, or even increase with sample size (as in a set-size effect, Anderson, 1967), their experiments were not designed to critically test these ideas.

We contend that strict theorizing is worthwhile (Fiedler, 2014, 2017), arguing that an averaging rule – though flexible in explaining different kinds of data (Dawes, 1979) – oversimplifies sample-based impression formation. Theoretical progress and a deeper understanding of how impressions depend on sample size can only be obtained when the trait sampling process is set apart from the cognitive trait-integration process. Our cognitive-ecological perspective on impression formation leads to several new insights, highlighting that intrapsychic processes can only be understood when the environmental sampling process is analyzed in the first place (Kutzner & Fiedler, 2015).

Our aim is not to study naturally occurring impression formation in a dynamic social context, as a function of perceivers' deliberate search strategies (Waggoner et al., 2009) or impressions in different stages of close relationships (Finkel et al., 2015). Rather, to answer the questions raised by Norton et al. (2007) and Ullrich et al. (2013), it is necessary to rule out the complexities of dynamic social interactions in an experimental design that relies on random sampling of traits. Within such an idealized experimental set-up, statistical sampling theory can be used to derive hypotheses about distinct causal influences on sample-based impression formation.

Functional Analysis of the Impression Formation Task

With this goal in mind, we extend the paradigm in several ways. First, we ask participants to provide impression judgments of multiple targets across trials, rather than only one judgment as in the preceding experiments. A sequential design does not only increase the overall data base and the reliability of empirical results. It also allows us to relate sample size to strength of impressions within participants, across targets described by varying numbers of traits.

Second, we manipulate the base-rates of positive and negative traits in different reference sets, from which the target traits are sampled. Across trials within participants, the positivity rate p_+ in the reference traits can take on four levels (.20 vs. .33 vs. .67 vs. .80), yielding repeated measures for positive ($p_+ > .5$) versus negative targets ($p_+ < .5$) and for moderate ($p_+ = .33$ or .67) versus extreme ($p_+ = .2$ or .8) targets. Such an overall flat distribution of impressions at all valence levels should prevent participants from anchoring their judgments in prior expectancies of moderately positive behavior, which is most common in the real world.

Third, the traits sampled from the four reference sets vary in diagnosticity. Negative and extreme traits are more diagnostic than positive and moderate traits (Koch et al., 2016; Peeters & Czapinski, 1990; Unkelbach et al., 2008). Regarding the "big two", negative traits related to morality (e.g., dishonesty) and positive traits related to ability (high expertise) are more diagnostic than positive morality traits (honesty) and negative ability traits (low expertise; cf. Fiske et al., 2007; Reeder and Brewer, 1979; Skowronski and Carlston, 1987). So, to analyze the impact of diagnosticity, we classify stimulus traits not only by valence and extremity but also by the big two, as referring to ability or morality.

Fourth, we introduce a crucial distinction between experimenter-determined and self-truncated sampling. Sample size is either set to $n = 2$, 4, or 8 in the *fixed-n condition*, or, in a *self-truncated sampling condition*, respondents can stop sampling at will. Finally, each participant in a *yoked-control condition* receives exactly the same stimulus samples as one yoked participant in the self-truncated condition. As we shall see in the next section, the impact of the number of traits on resulting impressions will be radically different in these three sampling conditions.

Last but not least, we include two different measures of the resulting impressions. In addition to participants' final subjective impression judgments, we include a measure of actuarial judgments (Dawes et al., 1989), defined as the average valence scale value of all traits in a sample. Actuarial judgments capture the stimulus samples that provide the input to the cognitive integration of sampled traits in the final impression judgments.

Implications Derived from Pertinent Theories

Within this enriched paradigm, a number of distinct theoretical implications can be tested. Let us first outline these basic implications, which may not all be common sense in social cognition research, before we lay out specific predictions and experimental methods.

Sampling Theory

To the extent that either positive or negative valence prevails the distribution of samples drawn from this universe will be skewed. The more common outcome will be overrepresented in the majority of samples, especially when sample size is small (Hadar & Fox, 2009; Hertwig & Pleskac, 2010). The deviation of the proportion of positive elements in the population p_+ from .5 reflects this skew and over-representation towards a positive ($p_+ > .5$) or negative ($p_+ < .5$) direction. When $p_+ = .8$, for instance, 64% of all samples of size $n = 2$ can be expected to obtain only positive traits (exceeding .8). Although the valence mean is unbiased, any strategy that does not purely reflect the mean but a tally of the most frequent outcome, or samples exceeding a high threshold (Fiedler & Kareev, 2006), will let small samples exaggerate the dominant outcome. What looks like a "less-is-more effect" in human impression formation may thus reflect a normal sampling effect. Neither Norton et al. (2007, 2011) nor Ullrich et al. (2013) mentioned the possibility that the dependence of trait valence on samples size may be already built into the sampling input, before the cognitive integration process started. The analysis of actuarial judgments (i.e., average valence scale values of sampled traits) provides a patent means of separating sampling effects from cognitive integration effects.

Impression Updating

Whereas the over-representation of the more prevalent outcome in small random samples alone may account for a less-is-more result, an opposite more-is-more effect derives from a theoretical analysis of impressions formed in a sequential updating process. Again, an increase in impression strength with increasing sample size can result from a fully unbiased cognitive process. To illustrate, assume that in the context of an experiment with many trials involving an overall balanced distribution of positive and negative targets an unbiased judge has to start from neutral expectations, consistent with the principle of insufficient reason (Bernoulli, 1713; Savage, 1954). The neutral starting impression is then updated sequentially by adjusting the growing impression to the incoming information. In an unbiased design, in which positive and negative traits are pretested to be of comparable strength, the expected upward and downward shift caused on average by positive or negative traits, respectively, should be the same. If a sample includes n_+ positive and n_- negative traits, the expected (unbiased) impression will be the neutral starting value plus n_+ upward minus n_- downward shifts. It is easy to see that the difference $(n_+ - n_-)$ increases with sample size: The expected difference for 10 traits drawn from a universe with $p_+ = .8$ is $(n_+ = 8)$ minus $(n_- = 2) = 6$; for a smaller sample of 5 traits the expected difference is only $(n_+ = 4)$ minus $(n_- = 1) = 3$.

Sequential updating thus renders impressions stronger (in the dominant direction) as sample size increases. Note that a repeated-updating mechanism, which produces a summation effect of $(n_+ - n_-)$ times the expected average shift, can be contrasted with an averaging model, which predicts a total updating effect of (n_+/n_-) times the average shift expected per trait, and which is therefore independent of absolute sample size[1]. Because actuarial judgments are defined as the average valence scale value of all sampled traits, any (non-average) more-is-more effect cannot be visible in the actuarial measure. It must originate in a cognitive trait integration process that deviates from the sampling input. A comparison of both measures may thus afford a useful tool to diagnose the origin of increasing impression strength with increasing sample size.

Diagnosticity

It seems obvious that the updating influence added by specific traits should depend on their diagnosticity. Negative traits are well known to be more diagnostic than positive traits (Fiske, 1980; Ito & Cacioppo, 2005; Skowronski & Carlston, 1987). Thus, when integrating multiple traits in an impression, the more diagnostic negative traits should have a stronger impact than positive traits. Despite their matched (symmetric) valence scale values, negative traits should have a stronger impact on the cognitive integration process than negative traits. Moreover, diagnosticity (in the sense of an enhanced updating influence) should be higher for extreme than for moderate traits, and for negative communion traits and positive competence traits than for positive communion and negative competence traits. As already noted by Fiske (1980), diagnostic impact is inversely related to expectedness or base-rates of occurrence. Rare and unexpected (extreme, negative, non-communal and high-agentive) traits are more diagnostic than common and expected (moderate, positive, communal and low-agentive) traits. Whether judgments are sensitive to diagnosticity can be tested by coding sampled traits in terms of all three sources of diagnosticity (valence, extremity, and the interaction of valence with the big two). Note, again, that diagnosticity cannot affect the (average) actuarial measure, because the scale values of positive and negative traits are of equal strength. Diagnosticity can only affect the post-sampling cognitive integration stage in the impression-formation process. The concept of diagnosticity therefore clearly differs from sampling mechanisms discussed in the last paragraph; it is expected to cause systematic asymmetries even in perfectly balanced (symmetric) experimental contexts.

[1] The discrepancy reflects the fact that the described updating algorithm starts from flat priors (i.e., neutral starting values), whereas the averaging algorithm is only sensitive to the valence of all stimuli, independent of any prior expectation. Because the summed valence is divided by n, averaging is independent of sample size.

Fixed Sample Size Versus Self-Truncated Sampling

In addition to the less-is-more effect due to overrepresentation of the more frequent valence in samples of restricted size, there is another type of a less is more effect, which can be very strong and visible in both human and actuarial judgments and which was never considered in previous impression research. This novel and neglected phenomenon cannot occur when sample size is manipulated experimentally at fixed levels of n. It only arises when sampling is self-truncated so that n depends on participants' own decisions to stop sampling as a sufficiently clear-cut impression has been reached. Self-truncated sampling only allows judges to decide when to stop a sample the ordering and contents of which is as randomized as in a fixed-n task. Thus, self-truncation does not allow judges to engage in active, self-determined search of specific stimulus items (as in Waggoner et al., 2009) or to influence the sequential order.

Yet, the seemingly minor difference between self-truncated random samples and experimenter-determined random samples regularly produces a profound primacy effect leading to a marked less-is-more effect. Computer-simulations demonstrate, indeed, that self-truncated sampling produces this phenomenon across a wide range of parameters (p_+ , and specific aggregation- and stopping rules; Prager, Harris, & Fiedler, 2017), and under various task conditions such as two-armed bandits (Fiedler et al., 2010) or multiple student sampling in a virtual classroom paradigm (Harris et al., 2017).

The underlying principle is easy to understand. When the first few randomly drawn traits happen to convey a clear-cut positive or negative impression, sampling will be truncated early. Exactly because strong initial information is what triggers early truncation, small sample size is naturally correlated with strong sampled contents (as assessed by the actuarial measure). Samples will be truncated later when the initially encountered evidence is mixed and equivocal. As a consequence, small samples are likely to evoke strong and confident judgments, and this should be evident not only in the final integrative impression judgments but also in the actuarial measure of average trait valence.

Moreover, because truncation is presumably sensitive to the diagnosticity of early traits rather than only to their valence scale value, the strength of the primacy effect that renders small samples stronger and more informative than large samples is further enhanced when the initial traits are high in diagnosticity. After all, self-truncating is contingent on judges' own growing impressions, which in turn depend on the diagnosticity of the so far encountered traits. Therefore, small truncated samples (informing strong and confident judgments) should be replete with diagnostic (negative, extreme etc.) traits.

Our analysis suggests that early truncation due to primacy of diagnostic and evaluatively consistent traits cannot be reduced to a hot-stove effect (Denrell & March, 2001) or the hedonic sampling rule discussed by Denrell (2005), Denrell and Le Mens (2007), and Fazio et al. (2004). Drawing on Thorndike's (1911) law of effect, these authors assume a hedonic preference to sample more from positive than from negative sources. As unpleasant samples are truncated, negative initial impressions cannot be corrected, thus creating a negativity bias. However, such a hedonic truncation rule cannot explain the breadth of the primacy effect. First, negative trait words are hardly of sufficient hedonic value to trigger abrupt truncation[2]. Second, hedonic sampling cannot account for valence-independent diagnosticity effects. And finally, hedonic sampling cannot explain the result in the following yoked-control condition.

[2] The negativity effect shown by Fazio et al. (2004) and simulated by Denrell (2005) disappears when sampling from negative sources is not abrupt enough (cf. Fiedler, Woellert, Tauber & Hess, 2013).

Brunswikian and Thurstonian Sampling

Drawing on prior work by Juslin and Olsson (1997), we introduce a distinction between Brunswikian sampling (of evaluative properties of stimulus targets in the environment) and Thurstonian sampling (of internal states or evaluative reactions in different judges, or within judges across different contexts). Brunswikian sampling of stimuli from an experimentally controlled environment is the focus of most familiar sampling theories (Fiedler, 2000; Stewart, Chater & Brown, 2006; Walasek & Stewart, 2015). The notion of Thurstonian sampling is less common and hardly ever considered, although it only highlights the obvious fact that judgments reflect a genuine interaction of external stimulus information and internal responses generated within human judges. According to the law of comparative judgments (Thurstone, 1927), the very same attitude target solicits different evaluative responses in different judges (or on different occasions within the same judge), and this inter-judge variance must be taken into account to understand differential target judgments.

To illustrate this genuine interaction of variance between stimuli and between judges, consider the self-truncated sampling condition. Participants in this condition stop the search process when their own internal evaluation is sufficiently clear-cut. Truncation is not exclusively determined by the objective properties (i.e., diagnosticity and valence consistency) of the traits sampled in the environment (i.e., Brunswikian sampling). It also depends on variation between different judges, whose responses to the same stimulus input can vary a lot. Because of memorized associations and self-generated thoughts, different judges can be more or less ready to stop and solicit a judgment from the available stimulus sample. This person-dependent variation in reacting to the very same (Brunswikian) stimulus input creates a crucial difference and asynchrony between the self-truncated search condition and a yoked control group of participants. Participants of the self-truncated condition truncated sample size not only because it appeared clear-cut based on its mere Brunswikian value, but also due to their individual and specific interpretation. Participants in the yoked-control condition however, cannot be expected to see the same samples as equally informative and hence to produce similarly strong judgments.

Empirical Predictions

As evident from this theoretical discussion, there can be no general answer to the question of whether an increasing number of n traits will produce stronger, weaker, or equally strong judgments. Under distinct conditions, different measures (actuarial, cognitive) of evaluative impression formation should be sensitive to different sample-size effects. The following predictions, derived on theoretical a-priori grounds, will be tested in three experiments:

We expect the impact of (Brunswikian) sampling error to be substantial, consistent with our cognitive-ecological approach. Brunswikian sampling error is the deviation of the sample average valence from the population's average it is drawn from. It represents valence information specific to the sample. In regression analyses conducted within individual participants across all trials, the resulting impression judgments should strongly depend on random variation of sampled traits (sample specificity), when the impact of systematic predictors (valence, extremity, sample size) is controlled for. This should be the case for all experiments, using fixed-n samples (Experiment 1), self-determined samples (Experiment 2a), and yoked controls (Experiment 2b). The impact of random sampling should be evident both in judges' final evaluations and in actuarial judgments (i.e., average scale values of sampled items).

Second, unsystematic sampling error should come along with systematic influences of the universe from which the stimulus samples are drawn. We predict extreme attributes (positivity rates of $p_+ = .20$ or $.80$) to trigger stronger judgments than moderate attributes (positivity rates of $p_+ = .33$ or $.67$), reflecting sensitivity to the parameters of the latent world. We also predict negative stimuli (positivity rates of $p_+ = .20$ or $.33$) to induce stronger judgments than positive attributes (positivity rates of $p_+ = .67$ or $.80$), due to enhanced diagnosticity of negative attributes (Gidron et al., 1993). We hasten to repeat that such valence asymmetry is not due to unequal scale values of negative and positive traits, which are controlled to be equal. Enhanced diagnosticity of negative traits must rather arise in the integration process, because negative attributes are more diagnostic, adding more independent and less redundant evidence than positive attributes, as specified in the density model (Unkelbach et al., 2008).

We also pursue the prediction concerning the Big Two that, orthogonal to general valence asymmetry, negative attributes should be more diagnostic in the morality domain whereas positive attributes should be more diagnostic in the ability domain (Reeder & Brewer, 1979; Skowronski & Carlston, 1987). To test this idea, we classify traits in a pilot study as referring to either morality or ability, and we include the variation in this sort of diagnosticity (in random sampling of traits) as a further predictor of individual judges' impression judgments.

Turning to the central research question concerning the impact of sample size, different predictions apply to three task settings. In the *fixed-n task*, impression judgments should either resemble the average scale value of all sampled attributes, showing little sample size effects if traits are low in diagnosticity. Or, judgment strength should increase with increasing sample size if diagnostic traits trigger noticeable updating. The latter condition might be met for extreme and negative traits, which have been shown to carry more diagnostic information than positive traits. Actuarial judgments should not exhibit such positive-negative asymmetry for the average scale values of positive and negative traits are matched. Any valence asymmetry must reflect post-sampling diagnosticity effects on the cognitive integration of traits.

In the *self-truncated search task*, in which judges can stop sampling when they feel to have gathered sufficient information, a primacy effect should render smaller samples more informative than larger samples. Judgments should increase with decreasing n to the extent that small (early truncated) samples entail a (joint) primacy effect in Brunswikian sampling (non-ambivalent and highly diagnostic initial traits) and in Thurstionian sampling (judges' internally generated responses to the stimulus input). If both truncation and trait integration are sensitive to unequal diagnosticity of positive and negative attributes, the less-is-more effect should be less visible for positive than for negative information.

Finally, judgments in the *yoked-control condition* rely on the same Brunswikian sampling input (captured by the actuarial measure) as judgments of yoked partners in the self-truncation condition. Yet, the primacy effect may be reduced or even eliminated in human impression judgments, due to Thurstionian sampling variation. Yoked-control judges' internal evaluations are not matched or synchronized with the evaluations experienced by judges in the self-truncated condition. As a consequence, impression judgments in the yoked-control condition should exhibit the primacy effect to a lesser degree. They should only exhibit the Brunswikian but not the Thurstonian component of the original judges' primacy effect.

Experiment 1

Methods

Participants and design. Forty-five participants (34 female) were recruited from the computerized subject pool “Studienportal” (software hroot: Bock, Nicklisch, & Baetge, 2012) at the University of Heidelberg. One participant whose response latencies to likeability questions were highly extended (more than three standard deviations above the participants’ median reaction time) was excluded. All three design factors – extremity and valence of stimulus traits and sample size – were varied within participants in a complete repeated measures design. All 2 (strong vs. moderate) x 2 (negative vs. positive valence) x 3 (sample size $n = 2$, 4, or 8) factor combinations were presented three times, yielding 2 x 2 x 3 x 3 = 36 judgment trials.

As it is impossible to estimate the effect size of a completely new design beforehand, we calculated the smallest effect size (post hoc) given the observed sample size of 44 participants. Setting type-I-error to $\alpha = .05$ and type-II-error to $\beta = .20$, the smallest effect size that could be reliably detected was $d = .38$. This effect size was slightly higher than the observed $d = .37$ for tests of the individual linear relationship between sample size and judgment strength against zero.

Materials and procedures. Fifty-seven trait adjectives of the Berlin Affective Word List – Reloaded (BAWL-R) by Võ, Conrad, Kuchinke, Urton, Hofmann, and Jacobs (2009) were selected for the present investigation. Four (overlapping) subsets of 30 traits out of these 57 BAWL-R adjectives served as reference sets (see appendix table A3), from which the stimulus samples of the 2 x 2 valence x extremity conditions were drawn.

The entire experiment was controlled by an interactive Java program. General introductory instructions were provided on the first screen, followed by a short demographic questionnaire and an agreement to participate conscientiously on the next screen. Then the procedure was explained in more detail. Participants were told that on every trial of a sequential judgment task they would be presented with several adjectives describing the traits of a target student, allegedly based on fellow students’ descriptions. Participants learned that they would be asked to judge the target person described on each trial on a likeability scale.

Each trial started with a blank screen after participants initiated a new trial by clicking on a corresponding button. The trait adjectives of the current sample appeared serially, one per 1000 ms, presented in the top center of the screen in black letters (font size 20 pt) on white ground. Only the last trait appeared in full contrast, whereas previous ones were dimmed to gray. When the stimulus sample was complete, a likeability scale appeared on the screen bottom. The poles of a 180 millimeter horizontal line were labelled “highly unlikeable” (in German: “starke Abneigung”) and “highly likeable” (“starke Zuneigung”). Participants could click on the continuous graphical scale to provide their evaluative impression. Afterwards, they were asked to indicate their subjective confidence on a five point scale, consisting of five discrete points labelled “very unsure”, “unsure”, “neutral”, “sure”, and “very sure”, respectively. The latencies of both responses (temporal difference between appearance and click on the scale) were recorded. Participants could then clear the screen and start the next trial by clicking any button. For each individual participant, the presentation order of the 36 trait samples was randomized, just as the ordering of traits within each sample. The experiment lasted for about ten minutes; it was the third out of five experiments conducted in a 60-minutes session.

Results

Relation of sample size to judgment strength. For a general index J of judgment strength, we computed the deviation of judgments from the scale midpoint in the direction of the predominant valence. That is, J scores were given a positive sign if $p_+ = .67$ or .80 but a negative sign if $p_+ = .20$ or .33, so that all measures were transformed to a scale reflecting judgment strength in the correct direction. The correlation $r(J, n)$ affords an appropriate measure of the linear relation between judgments strength (J) and sample size (n), which is the focus of the present and of the preceding investigations.

The average $r(J, n)$ computed within individual judges across all 36 trials (target persons) amounts to 0.05 ($SD = 0.14$), which is significantly different from zero, $t(43) = 2.44$, $p = .019$. The consensus rate of participants with a positive $r(J, n)$ is 64%. This overall figure is in line with the prediction of (slightly) increasing judgment strength with increasing sample size.

Closer inspection reveals, however, that this positive relation only holds for targets sampled from negative reference sets ($p_+ = .20$ or .33), average $r(J, n) = 0.11$ ($SD = 0.25$), $t(43) = 3.07$, $p = .004$, consensus rate 68%, but not for targets sampled from positive reference sets, average $r(J, n) > -.01$ ($SD = 0.26$), $t(43) = -0.12$, $p = 0.90$, consensus rate = 52%. We refrained from computing $r(J, n)$ for specific combinations of valence and extremity, because such correlations would be based on only 9 data pairs. However, the scatter plots in Figure 1 provide an overall picture of the strength of likeability judgments as a function of sample size n for all p_+ levels.

Systematic nature and sensitivity of impression judgments. To highlight the regularity and the sensitivity of impression judgments to all theoretically relevant aspects of the sampled traits, we conducted for each individual participant a regression analysis of the 36 judgments as a function of four predictors. In addition to the three orthogonal-design predictors extremity (extreme vs. moderate), valence (positive vs. negative), and sample size (2, 4, 8), we included sampling error as a fourth predictor. Sampling error is the deviation of the average scale value of sampled traits from the mean of the corresponding valence and extremity condition. As all four predictors are orthogonal – the first three by design and the last predictor stochastically – the standardized regression weight β for the sample size predictor must approximate the zero-order correlation $r(J, n)$. For the same reason, no interaction terms had to be included in the regression analyses. Table 1 exhibits mean β (SD in parentheses) averaged across all participants. The table also reveals the consensus proportions with which individual β-weights exhibit the same sign as the average judge, and t-statistics for all individual β-weights tested against zero.

Several strong and highly significant findings testify to the regularity of sample-based impression judgments. First, the individual β-weights for extremity are consistently positive, $t(43) = 6.59$, $d = .99$, $p < .001$. Judgments not only reflect the predominant valence but also the extremity of the predominant trend in the reference set. The consensus of positive β is 82%.

Second, the consensus (i.e. 100%) is maximal for the sampling error predictor, $t(43) = 27.96$, $d = 4.22$, $p < .001$, corroborating the judges' sensitivity to (even stochastic variation in) the input samples and, at the same time, highlighting the importance of taking the ecological sampling stage into account. Unsystematic noise due to ecological (Brunswikian) sampling error accounts for more systematic variance than any systematically manipulated influence factor. With regard to the third predictor, positive versus negative valence, it is evident, and actually not too surprising, that the vast majority of individual β-weights (82% consensus) are negative, indicating clearly stronger im-

Table 1. Regression Analyses of Likeability Judgments in Experiment 1 as a Function of Four Theoretically Relevant Predictors

Predictor	Mean β (SD)	Consensus	t value	df	p value
Extremity	.15 (.15)	82%	6.59	43	< .001
Valence	-.26 (.25)	82%	-6.85	43	< .001
Sample size	.08 (.13)	75%	4.11	43	< .001
Sampling error	.57 (.13)	100%	27.96	43	< .001

pression judgments triggered by negative than positive traits, $t(43) = -6.85$, $d = 1.03$, $p < .001$. Recall once more that such clear-cut valence asymmetry cannot be attributed to the pretested strength of negative and positive stimulus traits.

Finally, because $\beta = r$ for orthogonal predictors, the β-weights of the sample size predictor greatly resemble the aforementioned $r(J, n)$ results. A substantial majority of 75% positive β-weights, $t(43) = 4.11$, $d = .62$, $p < .001$, reflects a marked trend towards stronger judgments with increasing n. Theoretically, this central finding is not surprising; it could be expected on logical grounds if only some of the judges sometimes engage in updating of initially flat priors.

Diagnosticity. Analyses of trait diagnosticity support this account, showing that judgment strength only increases when added traits are diagnostic. To substantiate this point, we conducted another regression analysis within each judge, including four predictors that represent different aspects of diagnosticity: the extremity condition (extreme vs. moderate) along with three counts of traits per sample, (a) the frequency difference of positive minus negative traits in a sample, (b) the frequency difference of communion minus agency traits, and (c) the frequency of negative communion or positive agency traits minus the frequency of positive communion or negative agency traits. Note that (a) and (c), but not (b) are theoretically expected indices of diagnosticity. Because trait types (communion vs. agency) and valence (positive vs. negative) are hardly correlated across the stimulus materials (cf. Appendix), no interaction terms must be considered to interpret the β-weights.[3]

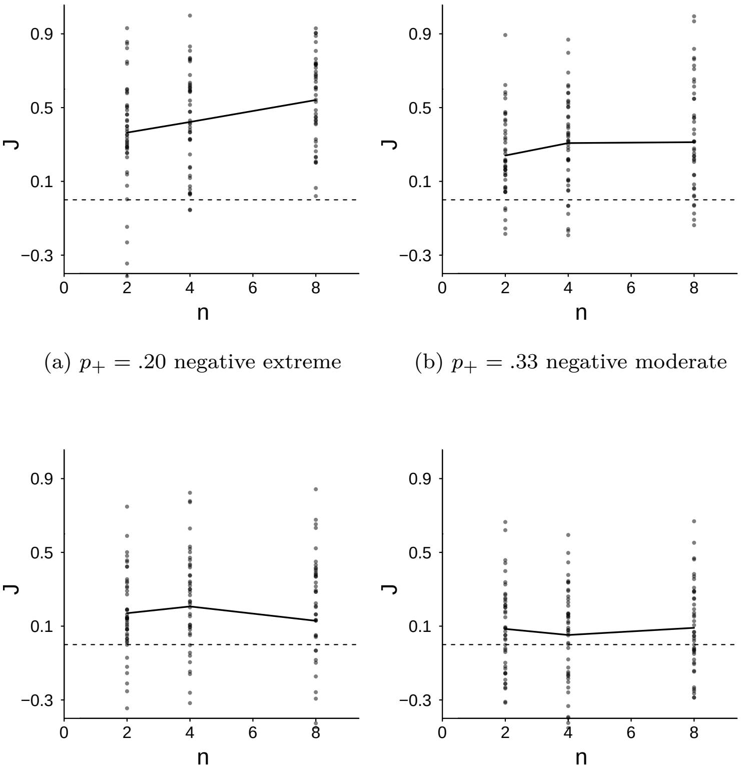
The mean regression weights in Table 2 corroborate the notion that likeability judgments are sensitive to (all measures of) diagnosticity. A consistently negative β-weight of the valence index validates the conclusion that negative traits inform stronger judgments than positive traits, $t(43) = -14.53$, $d = 4.38$, $p < .001$. Apparently, this trend, which holds for a majority of 98% of all judges, is particularly strong for the valence of the effectively sampled traits.

Interestingly, the other, conceptually independent measure of diagnosticity (i.e., the interaction of big two and valence) also contributes to predicting likeability strength. A positive β-sign for a majority of 86% judges, $t(43) = 7.59$, $d = 2.29$, $p < .001$, corroborates that negative communion and positive agency (i.e., diagnostic) traits trigger stronger impressions during the cognitive-integration stage than positive communion and negative agency.

The remaining index for the sheer frequency difference of communion and agency traits did not contribute, $t(43) = -1.01$, $d = -0.31$, $p = .317$; judgments did not depend on the relative number of communion versus agency traits. Note that the overall difference between communion and agency traits cannot account for the predictive value of the two other indices of diagnosticity.

The consistently positive β-weights of the extremity predictor, 80% consensus,

[3] Note also that the valence index in the present analysis (i.e., the different number of positive minus negative traits in a sample) is different from the valence predictor of the previous regression analysis underlying Table 1 (i.e., the p_+ baserate of positive traits in the reference set).



(a) $p_+ = .20$ negative extreme

(b) $p_+ = .33$ negative moderate

(c) $p_+ = .80$ positive extreme

(d) $p_+ = .67$ positive moderate

Figure 1. Strength of likeability judgments as a function of sample size (2, 4, or 8 traits) broken down by negative (upper plots) versus positive (lower plots) sign and extreme (left plots) versus moderate (right plots) strength of valence. Small grey dots represent individual judges' average judgments per condition; the solid black lines connect aggregated means across all judges. Note that likeability judgments of extremely negative targets ($p_+ = .20$) exhibit the strongest increase with sample size.

Table 2. Regression Analyses of Likeability Judgments in Experiment 1 Using Four Predictors Relevant to Assessing the Impact of Diagnosticity

	Real judgment criterion				
Predictor	Mean β (*SD*)	Consensus	*t* value	*df*	*p* value
Extremity	.15 (.17)	80%	5.73	43	$< .001$
Diagnosticity (valence)	-.34 (.16)	98%	-14.53	43	$< .001$
#Communion – # Agency	-.05 (.18)	61%	-2.09	43	.043
Diagnosticity (big two)	.18 (.16)	86%	7.59	43	$< .001$

$t(43) = 5.73$, $d = 1.73$, $p < .001$, replicate the results obtained in the first regression analysis. After all, extremity affords another index of diagnosticity. Thus, all three indices of diagnosticity provide strong convergent support for the contention that likeability judgments are sensitive to the evidence strength of a stimulus sample. This conclusion is consistent with the above interpretation of the influence of increasing sample size, which was strongest when trait samples were diagnostic in terms of two aspects, negativity and strength.

Confidence in the judgment. Recall that participants were asked to rate how confident they were in their likeability judgments after each trial. We refrain here from reporting all kinds of additional analyses that might be conducted with confidence ratings (e.g., different confidence-weighted likeability judgments), which all yielded sensible results. Suffice it to briefly mention the results of a regression analysis using confidence per se as criterion and the same four predictors as in Table 1 (reference sets' valence and extremity, sampling deviation from p_+, and sample size). First, positive β-weights for extremity (mean $\beta = .13$, $SD = .16$, consensus 82%, $t(43) = 5.65$, $d = .85$, $p < .001$) indicate that extreme samples induced higher confidence than moderate samples. Confidence also increased with increasing sampling error (i.e., fluctuation of the predominant valence in the sample's average valence; mean $\beta = .18$, $SD = .20$, consensus 75%, $t(43) = 5.76$, $d = .87$, $p < .001$). Thus, environmental sampling error not only had a strong impact on judgment strength but also on subjective confidence. Third, weak but significant β-weights of the valence predictor reflect somewhat higher confidence after judging negative than positive targets (mean $\beta = -.06$, $SD = .19$, consensus 59%, $t(43) = -2.06$, $d = .31$, $p = .046$). Finally, sample size did not affect the confidence of judgments (mean $\beta < .01$, $SD = .21$, consensus 55%, $t(43) = .31$, $d = .05$, $p = .760$). The latter finding speaks against the possibility that statistical education (e.g., the lesson that the reliability of a sample increases with n) may underlie the obtained impact of sample size on impression strength.

Discussion

To summarize, the first experiment supports the contention that sample-based impression judgments can be studied systematically, leading to distinctive and consistent findings. Likeability judgments are not only sensitive to the prevailing valence of the target's trait reference set but also to variation in degree of valence strength as well as two other indices of diagnosticity resulting from valence asymmetries in general and from the interaction of valence and the big-two (communion and agency), respectively. In addition to these influences on the cognitive integration process of the systematic target attributes defining the stimulus pool, the strongest predictor was the sampling error reflecting unsystematic variation in sampled trait valence. These (Brunswikian) sampling effects are antecedent to all cognitive processes.

With regard to the dependency of judgment strength on sample size, our findings highlight the benefits of a theory-driven research strategy. We reasoned that an updating process starting from flat priors would, if anything, produce stronger judgments with increasing sample size. This should only be the case if diagnostic traits render updating steps more likely. Consistent with this theoretical reasoning, we found the strength of likeability judgments to increase with increasing n, but only for negative traits – mostly for extreme negative traits. Sample size did not affect the strength of judgments in the positive domain. Marked valence asymmetry is consistent with many previous findings that testify to stronger impact of negative than positive traits on evaluative judgments (Kanouse & Hanson, 1987; Peeters & Czapinski, 1990), with a growing body of evidence on the density model (Alves et al., 2015; Unkelbach et al., 2008), and with linguistic evidence for enhanced diagnosticity of negative (compared to positive) social inferences (Rothbart & Park, 1986; Semin & Fiedler, 1992). Accordingly, updating should depend more on added negative traits than on added positive traits.

With regard to the previous debate between Norton et al. (2007) and Ullrich et al. (2013), our findings provide evidence concerning both positions, each under theoretically predictable conditions, which were masked by the designs used in the previous research. For the positive valence conditions that mostly resemble the restricted trait set used in these previous studies, our findings support the conclusion reached by Ullrich et al. (2013) that impression judgments are unaffected by sample size. In contrast, the positive influence of sample size on judgment strength is opposite to the less-is-more effect originally observed by Norton et al. (2007). However, the next experiment will demonstrate that a reverse ("less-is-more") relationship can be predicted theoretically when judges themselves can engage in self-truncated trait sampling, rather than receiving predetermined samples of fixed size. As we shall see, this shift to a more dynamic task setting, in which amount of information depends on judges' truncation decisions, will induce a strong negative relation between judgment strength and of sample size, that is, a clear-cut less-is-more effect that is predictable on theoretical grounds.

Experiment 2a (Self-Truncated Sampling)

Using the same basic stimulus materials and similar procedures as in Experiment 1, Experiment 2a also investigates sample-based likeability judgments, drawing on the same trait adjectives and the same three within-participants variables, strength, valence and samples size. However, the new set-up differs in one crucial aspect. Rather than receiving experimenter-controlled samples of fixed size, judges in this experiment are free to determine how many traits they want to see before they feel they can make a final judgment. Although common in real-life, self-truncated sampling leads to dramatic changes in theoretically expected sample-size effects. The reason is that, (Brunswikian) trait sampling is no longer independent of the (Thurstonian) variation in different judges' truncation decisions. Rather, n becomes a dependent variable that takes on different values for different levels of valence, extremity, and diagnosticity.

To the extent that judges are sensitive to the evidence of growing trait samples, a distinct primacy effect can be predicted, causing a diagnosticity advantage of small samples. On those trials, in which the first few randomly drawn traits happen to provide a clear-cut positive or negative impression, the search process will be truncated early and the resulting small samples will reflect decidedly strong impressions. This primacy effect (i.e., preponderance of one valence among the initial traits) will produce stronger judgments when early truncation renders sample size small.

Importantly, this should be evident not only in judges' actual likeability ratings but also in the average scale values of sampled traits (i.e., in actuarial judgments). Also, an analysis of the primacy effect affords a new intriguing test of diagnosticity in updating. The primacy effect should be most pronounced when negative (i.e., highly diagnostic) traits enhance the evidential value of the first few traits in a sample. Thus, given self-truncated sampling, the tendency for smaller samples to trigger stronger likeability judgments should be more pronounced when the dominant valence is negative rather than positive.

Method

The task set-up of Experiment 1 was used for Experiment 2a, except that the Java software was modified to allow participants to stop sampling when they felt appropriate. Instructions were adapted only slightly. Participants were asked to "only retrieve the number of traits [they] consider sufficient for making a judgment". The stimulus presentation procedure was modified accordingly: During information sampling, participants could either press the space bar to get an additional (randomly selected) trait or press the "Enter"-key (after at least retrieving one item) to terminate the sampling process and make a likeability judgment using the same scale format and subsequent confidence rating as in Experiment 1. During this stimulus-sampling process, shortened key-press instructions remained visible. Sample size was limited to a minimum of one and a maximum of 16 items.

Fifty-nine participants (46 females) took part in the experiment at the University of Heidelberg. Three participants were excluded because they invariantly sampled only one trait or the maximum of 16 traits. This time the post-hoc power analysis showed that sample size was large enough to reliably detect effects greater than $d = .34$ (given $\alpha = .05$ and $\beta = .20$). The observed effect size for the test of the individual linear correlations between sample size and judgment strength this time clearly exceeded this minimum effect size.

Results

Relation of sample size to judgment strength. Unlike Experiment 1, the linear correlations $r(J, n)$ in Experiment 2a between sample size n and judgment strength J, computed within each participant across all 36 trials (target persons), were clearly negative. Figure 2 portrays this strong reversal from a more-is-more to a less-is-more effect. The average overall $r(J, n)$, across all 36 trials, is –.21 ($SD = .24$), consensus 82%, $t(55) =$ ˘6.83, $p < .001$. However, again, the tendency of smaller samples to come along with stronger impression judgments is subject to a marked valence asymmetry. The correlation $r(J, n)$ between sample size n and judgment strength was clearly negative for negative reference sets ($r = -.38$, $SD = .27$, consensus 89%, $t(55) = -10.62$, $p < .001$). This was not the case for the positive reference sets ($r = .01$, $SD = .32$, consensus 46%, $t(55) = .27$, $p = .790$). The scatterplots in Figure 3 illustrate the valence asymmetry.

The opposite $r(J, n)$ relations obtained in Experiments 1 and 2a are significantly different from each other, across all 36 targets, $t(98) = -6.64$, $p < .001$, and of course for targets drawn from negative reference sets, $t(98) = -9.47$, $p < .001$, but not for targets drawn from positive reference sets, $t(98) = .27$, $p = .785$.

Systematic nature and regularity of impression judgments. Impression judg-

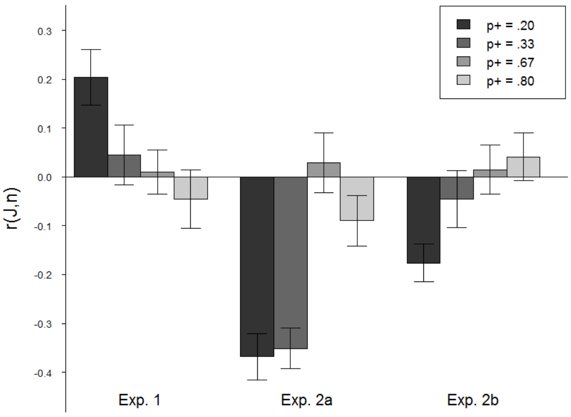
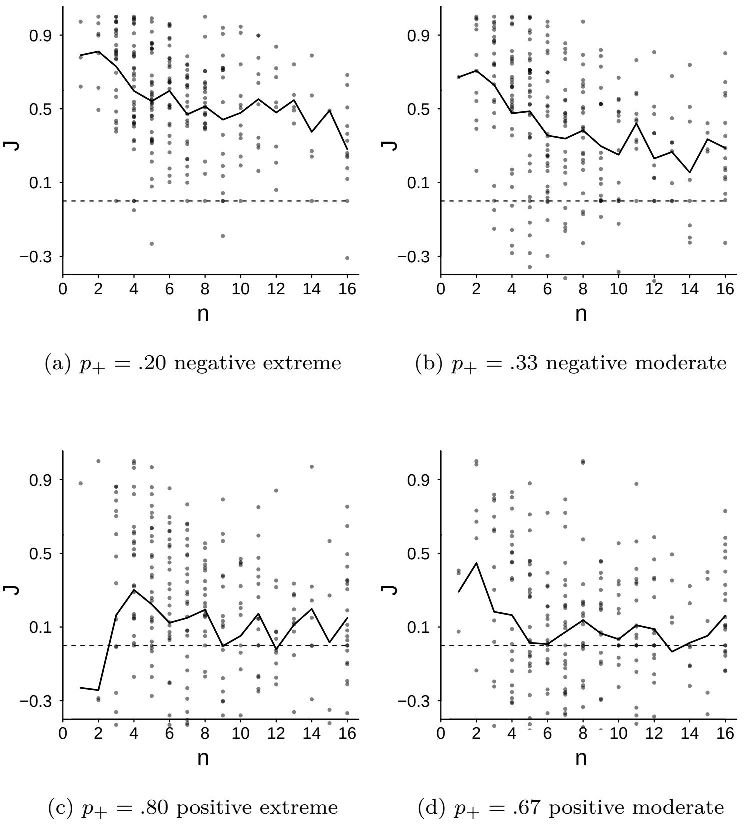


Figure 2. Mean linear correlation $r(J, n)$ between judgment strength and sample size of the current trait sample for all experiments, split up for the four reference sets (with different shades of grey representing different proportions of positive traits p_+ in the reference set). Error bars represent the standard error of the specific mean $r(J, n)$.



(a) $p_+ = .20$ negative extreme

(b) $p_+ = .33$ negative moderate

(c) $p_+ = .80$ positive extreme

(d) $p_+ = .67$ positive moderate

Figure 3. Experiment 2a: Strength of likeability judgments as a function of sample size, split up by reference sets with different p_+ proportions. Small grey dots represent individual judges' average judgments per condition; the solid black lines connect aggregated means across all judges. Note that likeability judgments of negative targets ($p_+ = .20$ or $.33$) exhibit the highest sensitivity to sample size.

Table 3. Predictor Intercorrelations for Experiment 2a

Predictor	Ref. set strength	Ref. set valence	Sample size
Ref. set valence	0		
Sample size	-.11*	.21*	
Sampling error	.01	-.07*	-.19*

Table 4. Regression Analyses of Experiment 2a, Using Judgment Strength as Criterion

Predictor	Mean β (SD)	Consensus	t value	df	p value
Reference set strength	.23 (.53)	66%	3.30	55	.002
Reference set valence	-.52 (.77)	79%	-5.00	55	$<$ .001
Sample size	-.06 (.27)	59%	-1.74	55	.087
Sampling error	.51 (.17)	100%	22.96	55	$<$ .001
Strength * sample size	-.11 (.51)	63%	-1.62	55	.111
Valence * sample size	.26 (.83)	61%	2.33	55	.023
Zero-order correlations with the extremity of the likeability judgment					
Reference set strength	.14 (.18)	77%	5.86	55	$<$ .001
Reference set valence	-.34 (.28)	89%	-9.14	55	$<$ .001
Sample size	-.21 (.24)	82%	-6.83	55	$<$ .001
Sampling error	.55 (.15)	100%	26.43	55	$<$ .001

ments based on self-truncated samples were again highly systematic and sensitive to all relevant information sources. We regressed each individual participant's 36 impression judgments (J) on the same four predictors as in Experiment 1: extremity, valence, sampling error (i.e., deviation from extremity and valence parameters), and sample size. Because self-determined sample size was no longer independent of other predictors (e.g., n could be different for predominantly positive and negative reference sets), we included the interactions between samples size and the extremity and valence predictors in the regression analyses. Note that the β weight of sample size deviates from $r(J, n)$ because correlated predictors (cf. Table 3) influence the relation between J and n.

A distinct pattern of regression results was obtained for a large majority of individual judges (cf. Table 4). Judgments were again sensitive to the extremity of p_+ ; the corresponding β was positive for 66% of the participants, $t(55) = 3.30$, $d = .44$, $p = .002$. Valence was also a significant predictor, $t(55) = -5.00$, $d = .67$, $p < .001$; β was negative for 79% of all participants; judgments were stronger for negative than for positive reference sets. By far the strongest predictor was again sampling error, the stochastic deviation of individual samples' average valence from the mean of the reference set. A positive sign of this predictor was obtained with a perfect consensus (100%), $t(55) = 22.96$, $d = 3.07$, $p < .001$.

Finally, regarding sample size, the β weight, mean $\beta =$ ˘.06, $t(55) =$ ˘1.74, $d = .23$, $p = .087$, was clearly weaker than the zero-order correlation $r =$ ˘.21 ($SD = .24$), consensus 82%, $t(55) =$ ˘6.83, $p < .001$. The stronger valence predictor apparently absorbed the predictive power of the redundant sample-size predictor ($r = .21$; cf. Table 3). A significant interaction of sample size and valence, $t(55) = 2.33$, $d = .31$, $p = .023$, means that early truncation decisions (yielding small n) were mainly due to the enhanced diagnosticity of negative traits. Although the interaction of sample size and extremity fell short of significance, $t(55) = -1.62$, $d = .22$, $p = .111$, its negative sign is also consistent with the notion that less-is-more effects tended to be stronger for diaganostic (i.e., extreme rather than moderate) targets (cf. Figure 3).

Diagnosticity. The mean β-weights of the other diagnosticity indices in Table 5 corroborate this. As in Experiment 1, extremity, $t(55) = 5.86$, $d = 1.57$, $p < .001$,

Table 5. Regression Analyses of Likeability Judgments in Experiment 2a Using Four Predictors Relevant to Assessing the Impact of Diagnosticity

	Real judgment criterion				
Predictor	Mean β (SD)	Consensus	t value	df	p value
Extremity	.14 (.18)	77%	5.86	55	< .001
Diagnosticity (valence)	-.30 (.17)	96%	-13.05	55	< .001
#Communion – # Agency	-.07 (.17)	70%	-2.89	55	.006
Diagnosticity (big two)	.14 (.18)	79%	5.87	55	< .001

positivity, $t(55) = -13.05$, $d = 3.49$, $p < .001$, and big-two diagnosticity, $t(55) = 5.87$, $d = 1.57$, $p < .001$, all contribute to predicting judgment strength. The frequency difference of communion minus agency traits also made a modest contribution; there was a significant trend towards stronger judgments as the relative number of agency traits increased, $t(55) = -2.89$, $d = -0.77$, $p = .006$.

Determinants of sample size. Examining self-determined sample sizes as a dependent measure is of interest in its own right. Individual regression analyses of n as a function of the extremity and the predominant valence of the target reference set showed that judges relied on smaller samples when reference sets were extreme (mean $n = 7.42$) than moderate (mean $n = 8.12$), mean $\beta = -.12$, $SD = .19$, consensus 75%, $t(55) =$ ˘4.66, $d = .62$, $p < .001$. Moreover, n was smaller for negative (mean $n = 7.13$) than for positive targets (mean $n = 8.41$), mean $\beta = .22$, $SD = .25$, consensus 80%, $t(55) =$ ˘6.66, $d = .89$, $p < .001$. Plausibly, a truncation threshold is reached faster when extremity and negativity render traits more diagnostic.

Frequency norms from the BAWL-R (Võ et al., 2009) allowed us to check on the notion that diagnosticity relates to unexpectedness (Fiske, 1980). Word frequency (according to BAWL-R-norms) was associated with self-determined sample size ($r = .19$, $SD = .22$, consensus 79%, $t(55) = 6.242$, $SD = .83$, $p < .001$). As low-frequency traits evoke surprise (non-familiarity), they have a stronger impact on truncation decisions than high-frequency traits, consistent with the hypothesized link between diagnosticity and unexpectedness.

Subjective confidence. An intriguing psychological concomitant of primacy effects due to early truncation is enhanced confidence. Initially strong tendencies in trait samples should not only lead to fast truncation of the search process. The resulting strong and congruent samples should also induce high degrees of confidence. Indeed, small sample size predicted confident likeability judgments, mean $\beta = -.32$, $SD = .41$, consensus 85%, $t(55) = -5.80$, $d = .78$, $p < .001$, thereby further emphasizing the (negative) relation of sample size to judgment strength.

Neither the extremity of target reference sets, mean $\beta = -.07$, $SD = .65$, consensus 58%, $t(55) = -.82$, $d = .11$, $p = .417$, nor the predominant valence, mean $\beta > -.01$, $SD = .73$, consensus 56%, $t(55) = -.06$, $d < .01$, $p = .953$, affected the confidence ratings. Interactions of sample size with extremity, mean $\beta = .15$, $SD = .67$, with consensus 65%, $t(55) = 1.65$, $d = .22$, $p = .104$, and with valence contributed little to predicting confidence either, mean $\beta < .01$, $SD = .83$, consensus 55%, $t(55) = .03$, $d < .01$, $p = .978$. These findings suggest that metacognitive feelings of confidence are not well-calibrated for diagnosticity. Only the β for sampling error was significant, mean $\beta = .16$, $SD = .18$, consensus 82%, $t(55) = 6.49$, $d = .88$, $p < .001$, reflecting higher subjective confidence when sampling error was consistent with the predominant tendency.

Discussion

Experiment 2 resembled Experiment 1 in many ways, corroborating the predictability of impression judgments from sensible psychological factors. With respect to sample size, the predictor of most interest, a seemingly small procedural change in task conditions produced a radical reversal. Whereas experimentally fixed sample sizes in Experiment 1 exhibited a "more-is-more" effect (i.e., stronger judgments with increasing n), self-truncated sample size in Experiment 2 had the reverse effect. Judgment strength decreased as n increased, due to a distinct primacy effect that renders small (i.e., early truncated) samples particularly informative.

Our theoretical analysis implies that participants in this experiment will not only rely on stimulus sampling in the environment. When judges can search as much information as they like, they can stop when clarity of the sample coincides with preparedness of the mind. However, the same (Brunswikian) samples (profiting from the same primacy effect) may not lead to the same strong and confident judgments when Thurstonian responses in different judges are not identical. For a simple and straightforward test of this intriguing notion, we conducted another experiment, in which exactly the same samples as in Experiment 2a were presented to new participants, who could however not themselves determine the moment of truncation.

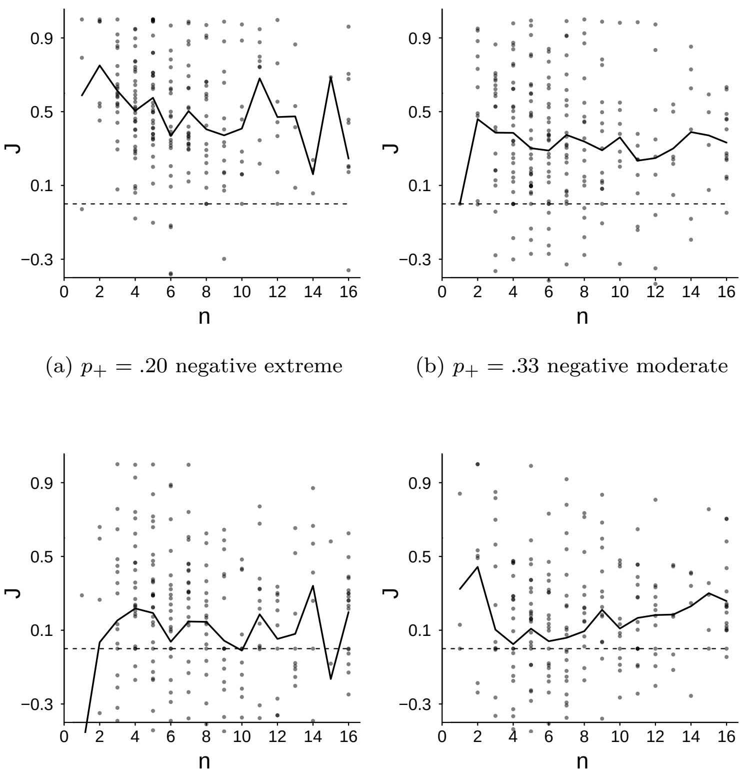
Experiment 2b (yoked controls)

Each participant of Experiment 2b was a yoked control of one participant in Experiment 2a, receiving exactly the same series of 36 samples. If the polarized judgments triggered by small samples can be explained in terms of Brunswikian sampling alone, a similarly strong "less-is-more" effect should be obtained. If, however Thurstonian sampling moderates the judgments, the same samples may not induce the same sample-size effect in Experiment 2b judges, who may not be in a critical state of mind when samples are truncated (by somebody else). In the latter case likeability judgments should be markedly diluted, due to misaligned Thurstonian activity.

Method

Fifty Heidelberg students (41 female) participated. One participant was excluded because of conspicuously long judgment latencies (> 3 standard deviations from the median latency). Six yoked control participants were missing in Experiment 2b, resulting in a slightly smaller of participants than in Experiment 2a (56). The minimum reliably detectable effect size in the post-hoc power analysis was $d = .36$ (given $\alpha = .05$ and $\beta = .20$) for the given sample size of 50. The observed effect size for the test of linear correlations between sample size and judgment strength exceeded this minimum effect size.

Materials and procedures. Instructions and procedures were the same as in Experiment 1; the stimulus materials were identical to Experiment 2a. Each participant of Experiment 2b received the same trait samples in the same order and size as the yoked partner in Experiment 2a, except for the active trait-sampling instructions. The yoked-control design was not mentioned.



(a) $p_+ = .20$ negative extreme

(b) $p_+ = .33$ negative moderate

(c) $p_+ = .80$ positive extreme

(d) $p_+ = .67$ positive moderate

Figure 4. Experiment 2b: Strength of likeability judgments as a function of sample size, split up by reference sets with different p_+ proportions. Small grey dots represent individual judges' average judgments per condition; the solid black lines connect aggregated means across all judges.

Results and Discussion

Relation of sample size to judgment strength. Comparison of the middle and right parts of the bar chart in Figure 2 shows that the average $r(J, n)$ across all 36 trials in Experiment 2b was clearly weaker, $r =$ ˇ.10, $SD = .22$, consensus 70%, $t(49) =$ ˇ3.39, $p = .001$, than in Experiment 2a above, $t(49) = 3.06$, $p = .003$, between experiments. The less-is-more effect, manifested in negative $r(J, n)$, was still stronger for targets drawn from negative references sets, ˇ.15 ($SD = .24$), although reduced in comparison to the latter experiment, $t(49) = 5.52$, $p < .001$. The correlation for positive reference sets was negligible, $r = .04$, $SD = .27$, consensus 52%, $t(49) = .94$, $p = .350$ (see Figure 4), and indifferent from Experiment 2a: $t(49) = 1.11$, $p = .273$ (see Figure 2).

Apparently, then, Thurstonian variation between judges contributed substantially to the enhanced less-is-more effect in Experiment 2a. Whereas judges in Experiment 2a truncated sampling when they were mentally ready to recognize a clear-cut trend in a growing sample, judges in Experiment 2b were exposed to the same sample, but at a moment when they may have not mentally prepared (i.e. their individual interpretation of the same sample did not lead to an equally clear-cut impression) to draw a judgment. Thurstonian sampling effects were not synchronized with the Brunswikian samples gathered from the environment.

Systematic nature and regularity of impression judgments. Though not in-

Table 6. Individual Regression Analyses of Judgment Strength in Experiment 2b

Predictor	Mean β (SD)	Consensus	t value	df	p value
Reference set strength	.21 (.54)	68%	2.76	49	.008
Reference set valence	-.28 (.62)	62%	-3.22	49	.002
Sample size	.15 (.31)	76%	3.43	49	.001
Sampling error	.56 (.18)	100%	22.49	49	$<$ .001
Strength * sample size	-.11 (.61)	60%	-1.26	49	.213
Valence * sample size	$<$ .01 (.64)	46%	.07	49	.943
Zero-order correlations with the extremity of the likeability judgment					
Reference set strength	.13 (.17)	70%	5.26	49	$<$ .001
Reference set valence	-.29 (.27)	84%	-7.38	49	$<$ .001
Sample size	-.10 (.22)	70%	-3.39	49	$<$ .001
Sampling error	.53 (.16)	100%	22.81	49	$<$ .001

Table 7. Differences in β-Weights Between Experiments 2a and 2b, Using Judgment Strength as Criterion

Predictor	$\Delta\beta$ (SD)	Consensus	t value	df	p value
Reference set strength	>-.01 (.77)	48%	-.03	49	.978
Reference set valence	.20 (.80)	64%	1.75	49	.087
Sample size	.21 (.38)	78%	3.88	49	$<$ .001
Sampling error	.05 (.18)	62%	1.93	49	.059
Strength * sample size	-.03 (.85)	52%	-.23	49	.822
Valence * sample size	-.21 (.88)	64%	-1.70	49	.095
Differences in zero-order correlations					
Reference set strength	-.03 (.14)	58%	-1.27	49	.209
Reference set valence	.05 (.40)	50%	.92	49	.361
Sample size	.11 (.25)	66%	3.06	49	.004
Sampling error	>-.01 (.17)	44%	-.37	49	.716

formed by one's "own sample", impression judgments in the yoked-control condition were as systematic and sensitive to all relevant regular manipulations as in both previous experiments. Table 6 summarizes the results of the pertinent within-participants regression analyses.

The extremity of the target reference set was again a significant predictor (consensus 69%, $t(49) = 2.76$, $d = .39$, $p = .008$); judgments tended to be stronger for targets sampled from extreme than from moderate reference sets. Sampling error (i.e., deviations of the average scale value of traits sampled from p_{+}) remained the strongest predictor (consensus 100%, $t(49) = 22.49$, $d = 3.18$, $p < .001$). Valence asymmetry was also replicated, with negative target sets producing stronger judgments than positive target sets (consensus 62%, $t(49) =$ ˘3.22, $d = .46$, $p = .002$).

Surprisingly, though, controlling for the other (correlated) predictors, the sample-size predictor now received a significant positive β-weight across all 36 judgments (mean $\beta = .15$, consensus 76%, $t(49) = 3.43$, $d = .49$, $p = .001$). The interactions of reference set strength with sample size (consensus 60%, $t(49) =$ ˘1.26, $d = .18$, $p = .213$) and of reference set valence with sample size (consensus 46%, $t(49) = .07$, $d < .01$, $p = .943$) did not show consistent results.

Comparisons across experiments revealed that the β-weight relating sample size to judgment strength (negative for the zero-order correlations) was significantly higher in Experiment 2b than in Experiment 2a (Table 7), consensus 78%, $t(49) = 3.88$, $d = .55$, $p = .001$, apparently due to Thurstonian sampling. No other predictor but sample size discriminated between the two matched experiments.

Diagnosticity. Analyses of diagnosticity indices were largely consistent with preceding experiments (cf. Table 8). Extremity, $t(48) = 5.06$, $d = 1.44$, $p < .001$, positivity,

Table 8. Regression Analyses of Likeability Judgments in Experiment 2b Using Four Predictors Relevant to Assessing the Impact of Diagnosticity

	Real judgment criterion				
Predictor	Mean β (SD)	Consensus	t value	df	p value
Extremity	.12 (.17)	69%	5.06	48	$<$ .001
Diagnosticity (valence)	-.35 (.17)	98%	-14.66	48	$<$ .001
#Communion – # Agency	-.05 (.18)	61%	-1.88	48	.066
Diagnosticity (big two)	.17 (.14)	88%	8.60	48	$<$ .001

$t(48) = -14.66$, $d = 4.19$, $p < .001$, and big-two diagnosticity, $t(48) = 8.60$, $d = 2.46$, $p < .001$, were significant predictors of judgment strength, but not the frequency difference of communion minus agency traits, $t(49) =$ ˘1.88, $d = -0.54$, $p = .066$.

Subjective confidence. Given the missing alignment of Brunswikian and Thurstonian sampling in Experiment 2b, the enhanced subjective confidence in small (early truncated) samples that we observed in Experiment 2a should be reduced. Indeed, smaller samples no longer led to a significant increase in confidence, mean $\beta =$ ˘.07, $SD = .36$, consensus 57%, $t(49) = -1.36$, $d = .19$, $p = .179$); compared to Experiment 2a, $t(49) = 2.34$, $d = .34$, $p = .023$. Neither extremity, mean $\beta = .14$, $SD = .65$, consensus 57%, $t(49) = 1.48$, $d = .21$, $p = .145$, nor valence of the reference set, mean $\beta = .16$, $SD = .76$, consensus 49%, $t(49) = 1.48$, $d = .21$, $p = .145$, were significant predictors of confidence in Experiment 2b. However, a strong influence of (trend-consistent) sampling error is manifested in positive regression weights (mean $\beta = .24$, $SD = .19$, consensus 91%, $t(49) = 8.88$, $d = 1.27$, $p < .001$). We did not obtain an interaction of extremity and sample size (mean $\beta = -.06$, $SD = .70$, consensus 53%, $t(49) = -.61$, $d = .09$, $p = .543$) but a noticeable interaction of valence and sample size (mean $\beta = -.25$, $SD = .83$, consensus 59%, $t(49) = -2.15$, $d = .31$, $p = .037$). The latter interaction shows that the negative relationship between sample size and confidence was slightly more pronounced for negative reference sets.

Actuarial Judgments

To set the sampling stage apart from the cognitive-integration process, and to further validate our theoretical interpretation of the relation between sample size and judgment strength, let us finally consider the actuarial "judgments", that is, the average valence scale values of all sampled traits (Dawes et al., 1989).

An analysis of these actuarial judgments J^* reveals that the less-is-more effect in Experiments 2a and 2b originates in the sampling stage, prior to the cognitive integration of the sampled traits. Thus, the linear relation $r(J^*, n)$ between n and actuarial judgments, which was approximately zero for Experiment 1 ($r =$ ˘.05, $SD = .12$, consensus 66%, $t(43) = -2.52$, $p = .016$), dropped to strong negative values in Experiment 2a and by definition also in Experiment 2b using the same samples in a the yoked-control design, $r =$ ˘.21, $SD = .22$, consensus 84%, $t(55) =$ ˘6.92, $p < .001$. This clearly shows that in Experiment 2a self-truncation already produced a less-is-more effect in the objective stimulus input: Smaller samples did contain systematically stronger (actuarial) evaluations than larger samples drawn from the same universe.

Note that deviations of actual human judgments J from the actuarial judgments J^* must reflect the complementary influence of the cognitive-integration stage. The mean linear relation $r(J, n)$ was significantly higher than $r(J^*, n)$ in Experiment 1, (.05 vs. ˘.05; consensus 73%, $t(43) = 4.42$, $p < .001$), consistent with an updating

Table 9. Regression Analyses of Experiments 1 and 2a, Using Actuarial Judgments as Criterion

Predictor	Mean β (SD)	Consensus	t value	df	p value
Experiment 1 (fixed sample sizes)					
Extremity	.29 (.03)	100%	63.73	43	$< .001$
Valence	.07 (.01)	100%	63.39	43	$< .001$
Sample size	$< .01$ ($< .01$)	52%	.65	43	.522
Sampling error	.95 (.05)	100%	132.84	43	$< .001$
Experiment 2a (self-truncated sample sizes)					
Extremity	.40 (.10)	100%	29.57	55	$< .001$
Valence	.09 (.02)	100%	29.30	55	$< .001$
Sample size	$< .01$ ($< .01$)	41%	.25	55	.807
Sampling error	.92 (.09)	100%	80.67	55	$< .001$
Extremity * Sample size	$> -.01$ (.01)	80%	-4.99	55	$< .001$
Valence * Sample size	$< .01$ (.01)	71%	3.72	55	$< .001$
Experiment 2a: zero-order correlations $r(J^*, n)$					
Extremity	.38 (.18)	98%	15.70	55	$< .001$
Valence	.02 (.16)	55%	.89	55	.379
Sample size	-.21 (.22)	84%	-6.92	55	$< .001$
Sampling error	.91 (.05)	100%	151.27	55	$< .001$

account of the more-is-more effect (i.e., the positive sign of the linear $r(J, n)$ relation). This (diagnosticity-dependent) updating process must take place during the cognitive-integration stage.

In Experiment 2a, the negative relations $r(J^*, n)$ and $r(J, n)$ are similarly strong, (-.21 and -.21; consensus 55%, $t(55) = -.27$, $p = .790$), indicating that human judgments are not regressive relative to the actuarial input. That is, human judgments exploit the full linear relationship between sample size and evaluation strength that is inherent in the actuarial input. We interpret this surprising lack of regression (cf. Fiedler & Unkelbach, 2014) in terms of the counteractive force of Thurstonian sampling effects. Judgments in Experiment 2a were not only made at the moment when the (Brunswikian) stimulus samples exhibit the strongest evaluation effects. They were also made at the very moment of truncation when (due to internal Thurstonian processes) participants are mentally ripe and ready to make a judgment. This synchronization of Brunswikian sampling of stimuli and Thurstonian states within judges may explain the strong (non-regressive) less-is-more effect in Experiment 2a.

Corroborating this interpretation, the negative relation $r(J, n)$ was markedly weaker than $r(J^*, n)$ in Experiment 2b, ($-.10$ vs. $-.20$; consensus 60%, $t(49) = 3.25$, $p = .002$), when yoked-control judges were not synchronized or mentally prepared to make a judgment at the moment of truncation (by judges of Experiment 2a). Although they received the same trait samples, Experiment 2b judges were less sensitive to the enhanced valence of small samples.

It is also interesting that the strong linear relationship $r(J^*, n)$ between sample size n and actuarial judgments J^* largely disappeared when (individual) regression analyses included other predictors, which apparently played a causal role in truncation decision. The pertinent results are summarized in Table 9, separately for actuarial judgments of Experiments 1 and 2.

Consistent with a cognitive-ecological sampling approach, the actuarial sampling input was sensitive to the extremity of traits in the reference set, (Exp. 1: mean $\beta = .29$; Exp. 2: mean $\beta = .40$), and to sampling error, (Exp. 1: mean $\beta = .95$; Exp. 2: mean $\beta = .92$), which was always the most powerful predictor. However, crucially, the valence factor did not produce the same strong negativity effect as in human judgments (Exp.

1: mean β = .07; Exp. 2a: mean β = .09). If anything the slightly positive β weights tended to point in the opposite direction. Obviously, because scale values of positive and negative trait stimuli were balanced, the negativity effect must originate from the cognitive integration stage and cannot affect actuarial measures.

Although actuarial judgments must be independent of sample size, (mean β < .01) and valence in Experiment 1, self-truncation in Experiment 2a (and 2b) creates distinct correlations between sample size and other variables related to truncation decisions. Due to higher diagnosticity of negative than positive traits, and of extreme than moderate traits, samples were truncated earlier when traits were negative and extreme (rather than positive and moderate), as evident in distinct influences of valence and extremity on actuarial judgments (cf. Table 9). Moreover, because valence and extremity triggered truncation and resulting sample size, the influence of sample size was absorbed by these two logically antecedent predictors and therefore only showed up in significant extremity x sample size and valence x sample size interactions.

To summarize, the less-is-more effect in Experiment 2 arose early in the sampling stage, prior to the cognitive integration process, when a few diagnostic traits created a strong primacy effect, encouraging early truncation of trait sampling. Whether the negative influence of sample size on valence strength carried over to actual human judgments depends on the alignment of Brunswikian and Thurstonian sampling. When Thurstonian and Brunswikian sampling effects were well aligned, as in Experiment 2a, then judges were mentally prepared for a judgment at the moment of truncation. So they exhibited a similarly strong less-is-more effect as the actuarial judgments. In contrast, when judges in Experiment 2b were provided with identical samples truncated by other participants, Thurstonian and Brunswikian sampling effects were not aligned, and judgments were therefore less sensitive to the primacy advantage.

General Discussion

Revisiting a debate in the present journal, we investigated the relation between impression formation and amount of information from a decidedly theory-driven perspective (Fiedler, 2017). The basic idea of strong cumulative science is to build testable hypotheses on firmly established empirical laws and logical principles. Strictly derived theoretical constraints can then be tested empirically.

Thus, with respect to the primary research question, the dependence of person impressions on the number of (randomly sampled) traits from the universe of all available information, it is essential to consider the logic underlying the principle of insufficient reason (Gilboa et al., 2009; Savage, 1954). In the context of a sequential task covering the full range of positive and negative trait proportions (i.e., p_+ varying from .20 to .80), this principle implies that there is no sufficient reason to expect on the next trial a target to represent a particular degree of positive or negative valence. Judges must be equally prepared to encounter targets of all valence levels. An unbiased impression formation process should therefore start from a neutral default expectation, and this neutral starting value should then be updated in the light of a series of target traits. Depending on whether a new trait is more positive or more negative than the current impression, it should cause an update in upward or downward direction, respectively. This updating process is largely independent from Bayesian principles of belief updating, they would however lead to highly convergent predictions.

Moreover, because the total updating effect is the sum of upward minus downward upgrades, it follows that the resulting degree of polarization (i.e., the strength of impression judgments in the direction of the more prevalent valence) should increase with sample size. Whether such a viable process hypothesis is actually borne out has to be tested, but our findings actually support the contention of such a more-is-more effect after sequential trait sampling, when multiple trials with a flat distribution of p_+ call for neutral “priors”.

Although it could not be predicted with certainty, the dependency of impression updating on the diagnosticity of sampled traits is well consistent with this sequential sampling algorithm. Three different versions of diagnosticity (related to valence, extremity, and the interaction of valence and the big two) lend convergent support to the role of diagnosticity. No more-is-more effect was obtained for non-diagnostic (positive) traits. In other words, we replicate the null findings reported by Ullrich et al. (2013) for targets described by mostly positive (not diagnostic) traits, even though for diagnostic (negative) traits we demonstrate a positive relation between sample size and impression strength.

With regard to the less-is-more effect reported by Norton et al. (2007), reflecting decreasing impression strength with increasing n, our sampling approach at least suggests a plausible account. Because their participants provided only a single impression judgment based on a sample drawn from a largely positive pool, and because positive traits are more common than negative traits in the real world, participants might have started from a default expectation of a clearly positive target. Further traits may have diluted this (too) positive starting impression, letting impressions regress to less positive levels.

Our multi-trial repeated measures design with a flat distribution of traits from all valence levels served to substantiate the principle of insufficient reason (making all valence levels equally likely) and thereby to rule out any uncontrolled vicissitudes of experiments resting on a single trait sample. In this regard, to be sure, our design is not directly comparable to the design used by Norton et al. (2007) and by Ullrich et al. (2013).

Yet, another hardly contestable principle predicts a less-is-more effect for a completely different reason than the one suggested by Norton et al. (2007). It was shown to arise in the self-truncated sampling condition, when sample size is not fixed as an independent variable but dependent on the judges’ own truncation decision. When samples are terminated at the very moment when the participant feels to have obtained a clear-cut impression, it follows that early truncated (i.e., small) samples must be informative and reflective of strong impressions, whereas less clear-cut trait sets should be truncated later, as manifested in larger sample size. This natural consequence of self-truncation can be shown in Monte-Carlo simulations (Prager et al., 2017) to generalize over wide ranges of the parameter space (n ranges, p_+ parameters, specific stopping rules etc.), but only when samples are self-truncated, not when n is under extraneous experimental control.

Note that the enhanced strength and clarity of small self-truncated samples constitutes a bias, resulting from the exploitation of a stochastic primacy effect, that is, the exploitation of samples that happen to provide a clear-cut picture at the beginning. In the context of this selective primacy effect, n is no longer under experimental control but has become a dependent variable (dependent on the stochastic distribution of trait valence and diagnosticity in early sampling stages). Yet, it is important to note that self-truncation, or control over the amount of information gathered, is typical for impression formation under many natural conditions. Moreover, the mechanism underlying the less-is-more effect of self-truncation can be generalized to other judgment and decision tasks, and it is not restricted to random sampling. On the contrary, the bias can be even more pronounced when non-random sampling allows people to render small samples decisive and informative.

The inclusion of an actuarial measure of sample contents allowed us to analyze the nature of small self-truncated samples, informing strong and confident judgments, more closely. They were replete with negative and diagnostic traits, which were clearly more likely to encourage early truncation than positive and non-diagnostic traits. Thus, unlike the impact of diagnosticity on sequential updating in the fixed-n condition, which only affected the cognitive integration stage but not the actuarial measure of sample contents, the impact of the primacy bias is strongly manifested in the actuarial measure. In other words, the primacy-dependent less-is-more effect is manifested in the sampling stage as well as the cognitive integration stage.

At the end, we are convinced that it is worth engaging in strict theorizing – beyond ad-hoc speculation and mere explicit announcement (pre-registration) of empirical hypotheses that could be easily replaced by their opposite. Strong cumulative science means to beware of the logical and theoretical constraints imposed on testable hypotheses. Had we only doubled (or tripled) the number of participants relative to prior studies, or had we only run exact replications, we could not have gained new insights about several non-intuitive phenomena.

This holds in particular for the seemingly exotic distinction between Brunswikian and Thurstonian sampling, which simply refers to a truism: Person impressions (much like other judgments and decisions) are not fully determined by the objective properties of stimuli sampled in the environment but by the internal cognitive and affective responses generated by different individuals (or across time and occasions within individuals). While the term "Brunswikian" refers to sampling in an uncertain environment, the term "Thurstonian" refers to complementary influences of different internally generating responses to the same stimulus samples.

As a consequence, we have seen that yoked control judges who were presented the same samples as participants in the self-truncated condition produced clearly weaker less-is-more effects. These divergent results reflect the different Thurstonian responses solicited by the same Brunswikian stimulus samples in the two experimental conditions. Apparently, for the yoked control participants, the truncated samples were not as well-suited for strong and confident impression judgments as for the original participants who could make their own truncation decisions. The reasons for divergence may be manifold. Waggoner et al. (2009) reasoned that passive yoked control participants (in a related but distinct paradigm) try to reach a quick and more superficial judgment. There is no support for such a difference in the present research, as the impression judgments in the yoked control condition are no less systematic and sophisticated (in terms of sensitivity to sampling error, extremity, and diagnosticity) than judgments in the other conditions. But the notion of "Thurstonian sampling" allows for many sources of variation between judges in different conditions. In any case, it highlights the genuine interaction of external stimulus variance and internal judgment variance.

The precisely stated algorithms and theoretically derived boundary conditions that have guided our research and that have been supported empirically can improve our understanding of diversely related phenomena, the different names of which prevented their theoretical integration. For instance the learning- and updating-principle underlying the more-is-more effect is at the heart of countless other demonstrations that information increases with the amount of stimulus observations: when judging oneself versus others (Moore & Healy, 2008), majorities or minorities (Fiedler & Wänke, 2009), ingroups versus outgroups (Brewer, 2007), or when testing focal versus alternative hypotheses (Fiedler & Wänke, 2009; Koriat et al., 1980). Conversely, the enhanced power of (selectively) truncated small samples may help to integrate such phenomena as choice overload in consumer decisions (Chernev et al., 2015), hot-stove effects and hedonically motivated sampling effects (Denrell & Le Mens, 2012; Fazio et al., 2004), or under-justification effects inspired by dissonance theory (Linder & Worchel, 1970).

No doubt, the present approach is not meant to provide a comprehensive account of the entirety of psychological influences on person impressions. Hardly any phenomenon in a multi-causal world is determined by a single causal process. Our demonstration that sampling rules applied to elementary traits alone can account for a whole pattern of clearly predictable influences does not preclude the operation of other influences. Thus, Norton et al. (2013) may be right to assume that the meaning of traits may change in the context of other traits, and the quantum theoretical model (Busemeyer et al., 2011) may offer a computational account of such meaning change. Likewise, order effects (Hogarth & Einhorn, 1992) may moderate the impact of traits, and valence asymmetries may be further elucidated in terms of neural system properties (Ito & Cacioppo, 2005).

The focus of the present research was on sampling influences rather than neural structures. Order effects were deliberately excluded and distinct trait interactions were minimized by random sampling from a sufficiently large universe of traits. But these are exactly the task constraints established in the preceding work that provided the starting point for the present research. In any case, with regard to the goal to base empirical research on clearly spelled out theoretical constraints (Fiedler, 2017), it seems fair to conclude that the study of social cognition can profit a lot from a cognitive-ecological sampling approach. In our own lab, we are now working on a computational (simulation) model supposed to provide a more comprehensive understanding of the generality and the confines of self-truncated sampling effects.

References

Alves, H., Unkelbach, C., Burghardt, J., Koch, A. S., Krüger, T., & Becker, V. D. (2015). A density explanation of valence asymmetries in recognition memory. *Memory & Cognition*, *43*(6), 896–909.

Anderson, N. H. (1967). Averaging model analysis of set-size effect in impression formation. *Journal of Experimental Psychology*, *75*(2), 158–165.

Anderson, N. H. (1981). *Foundations of information integration theory*. Academic Press.

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, *41*(3), 258–290.

Bernoulli, J. (1713). *Ars conjectandi: Opus posthumum*. Thurnisii.

Brewer, M. B. (2007). The social psychology of intergroup relations: Social categorization, ingroup bias, and outgroup prejudice. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles (2nd ed.).* (pp. 695–715). Guilford Press.

Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review*, *118*(2), 193–218.

Chernev, A., Böckenholt, U., & Goodman, J. (2015). Choice overload: A conceptual review and meta-analysis. *Journal of Consumer Psychology*, *25*(2), 333–358.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582.

Denrell, J., & Le Mens, G. (2012). Social judgments from adaptive samples. In J. I. Krueger (Ed.), *Social judgment and decision making* (pp. 151–169). Psychology Press.

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review*, *112*(4), 951–978.

Denrell, J., & Le Mens, G. (2007). Interdependent sampling and social influence. *Psychological Review*, *114*(2), 398–422.

Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science*, *12*(5), 523–538.

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology*, *87*(3), 293–311.

Fiedler, K. (2000). Beware of samples! a cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*(4), 659–676.

Fiedler, K. (2014). From intrapsychic to ecological theories in social psychology: Outlines of a functional theory approach. *European Journal of Social Psychology*, *44*(7), 657–670.

Fiedler, K. (2017). What constitutes strong psychological science? the (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science*, *12*(1), 46–61.

Fiedler, K., & Kareev, Y. (2006). Does decision quality (always) increase with the size of information samples? some vicissitudes in applying the law of large numbers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 883–903.

Fiedler, K., Renn, S.-Y., & Kareev, Y. (2010). Mood and judgments based on sequential sampling. *Journal of Behavioral Decision Making*, *23*(5), 483–495.

Fiedler, K., & Wänke, M. (2009). The cognitive-ecological approach to rationality in social psychology. *Social Cognition*, *27*(5), 699–732.

Fiedler, K., Wöllert, F., Tauber, B., & Heß, P. (2013). Applying sampling theories to attitude learning in a virtual school class environment. *Organizational Behavior and Human Decision Processes*, *122*(2), 222–231.

Finkel, E. J., Norton, M. I., Reis, H. T., Ariely, D., Caprariello, P. A., Eastwick, P. W., Frost, J. H., & Maniaci, M. R. (2015). When does familiarity promote versus undermine interpersonal attraction? a proposed integrative model from erstwhile adversaries. *Perspectives on Psychological Science*, *10*(1), 3–19.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889–906.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

Gidron, D., Koehler, D. J., & Tversky, A. (1993). Implicit quantification of personality traits. *Personality and Social Psychology Bulletin*, *19*(5), 594–604.

Gilboa, I., Postlewaite, A., & Schmeidler, D. (2009). Is it always rational to satisfy Savage's axioms? *Economics & Philosophy*, *25*(3), 285–296.

Hadar, L., & Fox, C. R. (2009). Information asymmetry in decision from description versus decision from experience. *Judgment and Decision Making*, *4*(4), 317–325.

Harris, C., Hess, P., & Fiedler, K. (2017). *Self-truncated and externally determined sampling effects in the simulated classroom.* [Unpublished manuscript.].

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, *115*(2), 225–237.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55.

Ito, T. A., & Cacioppo, J. T. (2005). Variations on a human universal: Individual differences in positivity offset and negativity bias. *Cognition and Emotion*, *19*(1), 1–26.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, *104*(2), 344–366.

Kanouse, D. E., & Hanson, L. J. (1987). Negativity in evaluations. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 47–62). Erlbaum.

Koch, A., Alves, H., Krüger, T., & Unkelbach, C. (2016). A general valence asymmetry in similarity: Good is more alike than bad. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *42*(8), 1171–1192.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 107–118.

Kutzner, F., & Fiedler, K. (2015). Information sampling and reasoning biases: Implications for research in judgment and decision making. In G. Keren & G. Wu (Eds.), *The wiley blackwell handbook of judgment and decision making* (pp. 380–403). Wiley.

Linder, D. E., & Worchel, S. (1970). Opinion change as a result of effortfully drawing a counterattitudinal conclusion. *Journal of Experimental Social Psychology*, *6*(4), 432–448.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.

Norton, M. I., Frost, J. H., & Ariely, D. (2007). Less is more: The lure of ambiguity, or why familiarity breeds contempt. *Journal of Personality and Social Psychology*, *92*(1), 97–105.

Norton, M. I., Frost, J. H., & Ariely, D. (2011). Does familiarity breed contempt or liking? comment on Reis, Maniaci, Caprariello, Eastwick, and Finkel (2011). *Journal of Personality and Social Psychology*, *101*(3), 571–574.

Norton, M. I., Frost, J. H., & Ariely, D. (2013). Less is often more, but not always: Additional evidence that familiarity breeds contempt and a call for future research. *Journal of Personality and Social Psychology*, *105*(6), 921–923.

Peeters, G., & Czapinski, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology*, *1*(1), 33–60.

Prager, J., Harris, C., & Fiedler, K. (2017). *Systematic sampling biases arising from self-truncated stimulus presentation: A simulation study.* [Unpublished manuscript.].

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, *86*(1), 61–79.

Rothbart, M., & Park, B. (1986). On the confirmability and disconfirmability of trait concepts. *Journal of Personality and Social Psychology*, *50*(1), 131–142.

Savage, L. J. (1954). *The foundations of statistics.* Wiley.

Semin, G. R., & Fiedler, K. (1992). *Language, interaction and social cognition.* Sage Publications, Inc.

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology*, *52*(4), 689–699.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26.

Thorndike, E. L. (1911). *Animal intelligence. experimental studies.* Macmillan & Company.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286.

Ullrich, J., Krueger, J. I., Brod, A., & Groschupf, F. (2013). More is not less: Greater information quantity does not diminish liking. *Journal of Personality and Social Psychology*, *105*(6), 909–920.

Unkelbach, C., Fiedler, K., Bayer, M., Stegmüller, M., & Danner, D. (2008). Why positive information is processed faster: The density hypothesis. *Journal of Personality and Social Psychology*, *95*(1), 36–49.

Võ, M. L. H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin affective word list reloaded (BAWL-R). *Behavior Research Methods*, *41*(2), 534–538.

Waggoner, A. S., Smith, E. R., & Collins, E. C. (2009). Person perception by active versus passive perceivers. *Journal of Experimental Social Psychology*, *45*(4), 1028–1031.

Walasek, L., & Stewart, N. (2015). How to make loss aversion disappear and reverse: Tests of the decision by sampling origin of loss aversion. *Journal of Experimental Psychology: General*, *144*(1), 7–11.

Appendix A. Construction of the Four Trait Reference Sets

As mentioned above, fifty-seven trait adjectives of the Berlin Affective Word List – Reloaded (BAWL-R) by Võ et al. (2009) served as a pool of potential stimuli. For the experiments four different, but overlapping trait reference sets had to be drawn from the overall pool of stimuli.

Table A1. Frequency Distributions Underlying the Generation of the Four Trait Reference Sets Forming the Stimulus Material of All Three Experiments

	BAWL-R valence interval					
Reference set	-3 : -2	-2 : -1	-1 : 0	0 : 1	1 : 2	2 : 3
Extremely negative	3	12	9	4	1	1
Moderately negative	2	9	9	7	2	1
Moderately positive	1	2	7	9	9	2
Extremely positive	1	1	4	9	12	3

The first step in this procedure was to determine frequency distributions of resulting reference sets. The four distributions can be considered to consist of two pairs symmetric in valence. One moderate and one extreme distribution was formed and exactly mirrored towards the other direction of valence. Binomial distributions of $n = 5$ and p-parameters of .2, .4, .6, and .8 served as orientation. Note, that binomial distributions (equal n and p-parameters p and p^*) of $p^* = 1 - p$ are axisymmetric to the center. As $n = 5$ suggests, we split the scale of the BAWL-R valence norms (ranging from -3 to +3) into 6 intervals of equal range (1). Setting population size of each reference set to 30, the four frequency distributions presented in Table A1 resulted.

In a second step, the described frequency distributions had to be filled with stimuli. Therefore the 57 traits were split into the six intervals depending on their BAWL-R valence. Thus, 6 distinct urns of homogenous valence resulted. Next, the number of traits – determined by the previously formed distributions – were drawn randomly from each urn. All items of the urns were replaced after the stimuli for one reference set were drawn, so that within reference set items do not overlap, whereas one item might be part of different reference sets. Manifest statistics of the drawn reference sets are shown in Table A2.

Table A2. Statistics for Valence Scores of the Four Reference Sets

Reference set	Mean	p_+	*SD*	Skewness
Extremely negative	-.78	.20	1.10	.73
Moderately negative	-.40	.33	1.20	.43
Moderately positive	.50	.67	1.21	-.41
Extremely positive	.85	.80	1.20	-1.07

As can be seen from the current table, reference sets were almost, but not perfectly symmetric around neutral valence. Not only proportions of positive valent items p_+ varied systematically and symmetrically, but also the mean valence and skewness of the trait valences. Therefore the two extreme reference sets' p_+ deviated more from .5, but they also showed more deviating means and skewness (from 0) compared to the moderate reference sets. The finally selected trait reference sets can be seen in Table A3.

Table A3. Selected Trait Adjectives Together With Their English Translation and Their BAWL-R Valence (Scaled From Very Negative [-3] to Very Positive [3]; Võ et al., 2009)

Original German	English translation	BAWL-R valence	$p_+ = .20$	$p_+ = .33$	$p_+ = .67$	$p_+ = .80$
herzlos	heartless	-2.5	1	0	0	1
verlogen	mendacious	-2.3	1	1	0	0
humorlos	humorless	-2.1	1	1	1	0
boshaft	mischievous	-1.9	1	0	0	1
gemein	mean	-1.9	1	1	0	0
launisch	moody	-1.9	1	1	1	0
gierig	greedy	-1.5	1	1	0	0
labil	labile	-1.5	1	1	0	0
stur	obstinate	-1.5	1	1	0	0
mutlos	despondent	-1.4	1	0	0	0
primitiv	primitive	-1.4	1	1	0	0
altklug	precocious	-1.1	1	1	1	0
derb	coarse	-1.1	1	0	0	0
eitel	vain	-1.1	1	1	0	0
passiv	passive	-1.1	1	1	0	0
wortkarg	taciturn	-0.9	1	1	1	0
laut	noisy	-0.8	1	1	1	1
naiv	naive	-0.8	1	1	1	1
defensiv	defensive	-0.6	1	1	1	0
unnahbar	inapproachable	-0.6	1	1	1	0
listig	cunning	-0.4	1	1	1	1
albern	foolish	-0.2	1	1	1	0
forsch	outspoken	-0.2	1	1	0	1
redselig	talkative	-0.1	1	1	0	0
stoisch	stoical	0.1	0	1	1	1
sparsam	thrifty	0.2	1	0	1	0
ruhig	calm	0.6	0	1	1	1
sachlich	factual	0.6	1	1	1	1
still	silent	0.6	1	1	1	1
verwegen	audacious	0.6	1	1	1	1
eifrig	eager	0.8	0	1	1	1
sensibel	sensitive	0.8	0	1	1	1
strebsam	ambitious	0.9	0	0	1	1
vornehm	genteel	0.9	0	0	0	1
liberal	liberal	1.3	0	0	0	1
sanft	gentle	1.3	0	0	1	1
spontan	spontaneous	1.4	1	0	1	1
aktiv	active	1.6	0	0	1	1
pfiffig	smart	1.6	0	0	1	1
schlau	shrewd	1.6	0	0	1	1
flexibel	flexible	1.7	0	0	1	1
munter	alert	1.7	0	0	0	1
tapfer	courageous	1.7	0	0	0	1
heiter	cheerful	1.8	0	1	1	1
nett	nice	1.8	0	1	1	1
taktvoll	considerate	1.9	0	0	1	1
loyal	loyal	2.1	0	0	0	1
ehrlich	honest	2.2	1	0	0	0
mutig	brave	2.2	0	0	1	1
treu	faithful	2.2	0	1	1	0
kreativ	creative	2.6	0	0	0	1

Note. The columns on the right hand side indicate the usage (1) or nonusage (0) of the trait for each of the four reference sets indicated by their proportion of positive valent traits.

Appendix B

Forming Impressions from Self-Truncated Samples of Traits – Interplay of Thurstonian and Brunswikian Sampling Effects.

Forming Impressions From Self-Truncated Samples of Traits – Interplay of Thurstonian and Brunswikian Sampling Effects

Johannes Prager and Klaus Fiedler

Heidelberg University

ABSTRACT

Consistent with sampling theories in judgment and decision research, impression judgments depend in distinct ways on the number of traits drawn randomly from a population of target person traits. When sample size is determined externally by the experimenter, the sensitivity of resulting impression judgments to the prevailing (positive or negative) valence increases with the number of traits. In contrast, sensitivity is negatively related to sample size (more extreme judgments for smaller samples) when sampling is self-truncated. Building on previous findings by Prager, Krueger, and Fiedler (2018), two new experiments corroborate the judgment pattern for self-truncated sampling and elaborate on the distinction of Brunswikian sampling (of stimuli in the environment) and Thurstonian sampling (of states within the judge's mind). Thurstonian sampling effects were evident in depolarized (regressive) judgments by yoked control participants provided with exactly the same trait samples as original judges, who could truncate sampling when they felt ready for a judgment. Experiment 1 included two kinds of yoked controls, receiving trait samples truncated in a previous stage either by themselves or by other judges, distinguishing between temporal and interpersonal sources of Thurstonian sampling variance. As expected, self-yoking yielded less regressive shrinkage than other-yoking. Experiment 2 provided convergent results with yoked controls manipulated within participants, dealing with higher dispersion of impressions on self-truncated samples (Thurstone, 1927). Across both experiments, individual impression judgments were highly predictable from theoretically meaningful parameters: expected valence in the population, sampling error, sample size and different indices of trait diagnosticity.

KEYWORDS

Thurstonian sampling, yoked controls design, diagnosticity, self-truncated sampling

Supplemental materials: https://doi.org/10.1037/pspa0000274.supp

Forming impressions from restricted information samples is a social-cognitive task that one frequently encounters in everyday life. Imagine, for example, a teacher asking knowledge questions to students, a consumer reading product reviews on the internet, or a personnel manager interviewing job candidates. What the examples have in common is a judge (teacher, customer, or personnel manager), whose task is to gauge a target

Johannes Prager, Department of Psychology, Heidelberg University, johannes.prager@psychologie.uni-heidelberg.de

Klaus Fiedler, Department of Psychology, Heidelberg University

The work underlying the present article was supported by a Grant provided by the Deutsche Forschungsgemeinschaft (Fi 294/29-1) to Klaus Fiedler. Helpful comments by Gaël Le Mens and Linda McCaughey on a draft of this article are gratefully acknowledged.

entity (student, product, job applicant) regarding a latent attribute (ability, quality, future job performance). This attribute is often not amenable to direct observation but needs to be inferred indirectly from restricted samples of raw observations. Judgments (and consequent decisions) entail inferences of latent attributes from restricted and often rather small and incomplete samples of relevant stimulus observations. Person impression judgments, in particular, are integrative verbal or numerical summaries of inferences drawn from samples of traits or behaviors.

One chief determinant of judgments and decisions is the amount of information, or the size of the sample it is based on. Assuming unbiased random sampling and sample size being an independent variable rather than dependent on the information content, the reliability and sensitivity of sampled evidence will normally increase with increasing sample size. According to the "law of large numbers" (Bernoulli, 1713) , as samples size increases, sample estimates approximate the true properties of the universe from which the sample is drawn. This widely known advantage of increasing sample size is not only evident in the reliability of scientific research but also in inference tasks of judgment and decision making (e.g. Peterson Beach, 1967), the wisdom of crowds (Galton, 1907) or more generally in the positive slope of learning curves. It is well known that the reliability of a test increases with the number of items (Spearman, 1910) and it is commonly supposed that the quality of expert judgments and advice increases with the amount of experience.

But while this is common sense, it seems worthwhile explicating the underlying logic. Statistical theory tells us that sample means afford unbiased estimates of the latent population parameter or its expected value, independently of sample size. However, while there is no reason to expect the mean or central tendency of a small sample to be more biased relative to a large sample, the crucial difference lies in the dispersion of the sampling distribution. Small samples are more dispersed and deviate more from the population means than larger samples. As a consequence of this basic inaccuracy, small samples are prone to misrepresent true population tendencies. However, the same inaccuracy that often obscures population properties may also exaggerate and over-accentuate existing trends (Hadar & Fox, 2009; Hertwig & Pleskac, 2010). As a consequence, smaller samples may, under distinct conditions, provide a particularly clear-cut and conflict-free picture of a true latent trend.

Imagine, for example, a student with a true probability $p = .8$ of responding correctly to arithmetic problems in a math lesson (i.e., the expected probability of responding correctly is 80%). A small sample of only $n = 2$ responses will most likely exhibit a correctness proportion higher than 80%; the most frequent sample proportion obtained at a rate of $.8^2 = .64$ is indeed 100%. Such an extreme proportion (that exceeds the expected value of 80%) is much less likely for a larger sample of, say, $n = 10$, (i.e., $.8^{10} = .11$). More generally, because sampling distributions are skewed, especially for small n, the smaller samples are more likely to exaggerate distribution trends, despite a constant expected value for every sample size (De Finetti, 1937; Hadar & Fox, 2009; Hertwig & Pleskac, 2010; Kareev, 2000).

Crucially, we implicitly assumed that sample size is a sample characteristic independent of its content, especially of the described initial instances of exaggeratedly clear outcomes. But what happens when we abandon this assumption and, instead cede control of the amount of sampled information to the individuals themselves, thus allowing sample size to become dependent on observed sampled information and the judging individual's strategy? That case is indeed quite representative of sampling in many natural settings. Teachers, for instance, tend to stop asking a particular student questions when their impression of the student's competence is sufficiently clear. Similarly, consumers decide themselves when to stop sampling and make a choice between products. Personnel managers conclude a job interview when they feel sufficiently informed to make a selection decision. In all these cases, self-truncation has seemingly paradoxical implications.

Self-truncation typically creates a negative correlation between sample size and strength of evidence in the sample, making existing trends more visible in small (early truncated) than in large (late truncated) samples. Crucial to understanding this curious phenomenon is the question of when sampling is truncated. Imagine an individual involved in a person judgment task. Their impression of a target person might already be clear after a few observations if all stimuli in an initial set convey a consistent picture. This high convergence of evidence will result in high confidence in the stability of the current impression. In this case, sampling will likely stop early, and the resulting small samples will result in clear-cut and confident impressions. Alternatively, a growing sample may as well be ambivalent and indeterminate at a small sample size. Impressions will seem confusing and uncertain and will likely motivate further information search due to too much variation and conflicting evidence in the initial sample. In the latter case, sampling is likely to continue and, in many cases, the resulting large sample remains equivocal and conflict-prone, leading to weaker impression judgments.

The strong evidence found in small samples is exactly the reason why they remained that small. Conversely, that large samples often carry weaker evidence was the reason why they were not stopped at an earlier stage. At first glance, the negative relation of impression strength and sample size seems to reverse common intuition that large samples are more reliable. As we examine the phenomenon more closely, we see however that all assumptions are perfectly compatible with sample-size related statistical principles, such as the law of large numbers.

We already implied instances of clear-cut impressions as reasons for judging individuals' decision to truncate an unfolding information sample. Although we refrain from propagating exclusive assumptions on invariant stopping mechanisms, the decision to truncate sampling flexibly follows the clarity in information, the diagnostic (informative) value of the sampled content and the individual's preparedness to provide an impression. In principle, it makes sense for truncation decisions to follow one or both of two dimensions: variance between targets and/or invariance within a focal target. If a task calls for a choice between two or more options, we may stop sampling when the differential evidence gathered for one and against alternative options is strong enough. However, graded evaluations from samples of information on a single target person are of greater relevance to the present research than the contrast between optional targets (as in personnel selection). In the present estimation and impression context, we may stop sampling when the information settles on a clear-cut and conflict-free impression that remains stable when new stimuli are added.

Brunswikian and Thurstonian Sampling

Up to this point, we have mainly discussed stimulus sampling that takes place in the environment, but we have disregarded the sampling of internal events that takes place within the individual's mind. Suppose that one person has truncated sampling at a point where they were sufficiently confident to make a judgment. Then this very same sample is presented to another individual, making them a *yoked control* pair. From an ecological or statistical sampling perspective the information observed by both individuals is equivalent. Both are exposed to identical target information, which ought to result in equivalent impression judgments. But are judgments informed by one's own sampling process really psychologically equivalent to judgments informed by samples passed on from another person?

Adopting a pair of terms coined by Juslin and Olsson (1997), Prager et al. (2018) proposed a distinction between *Brunswikian* and *Thurstonian* sampling, suggesting a negative answer to the question of whether the information input is equivalent for yoked controls and for persons who have themselves truncated a sample. For two independent observers of the same sample, the Brunswikian sampling input of stimuli provided by the environment is indeed identical. Brunswikian sampling is constrained by ecological factors alone, such as the true properties of the constitutive entity and the rate and strength of sampling error. Thurstonian sampling, in contrast, broadens the notion of sampling to a process that is not confined to information provided by the environment. It also includes internal sampling of the individual's states of mind, decision weights, encoding foci, attention and interference in observing information. Thurstonian sampling covers a variety of oscillating cognitive activities.

Diagnosticity and Expectedness

One particularly important variable that is sensitive to Thurstonian sampling is diagnosticity. In the context of impression formation, diagnosticity is crucial for the judge's metacognitive evaluation of the stimulus input and for the decision when to truncate. Diagnosticity of sampled information can be conceived as the ease with which the individual can locate the impression target on a relevant judgment scale, given the current sample of information (Skowronski & Carlston, 1987).

Imagine for example an impression-formation process in the context of a job interview or an episode of getting acquainted. Typically, all applicants report their professional success, interest in the company and present themselves as socially competent. Observing such reports in a job interview hardly differentiates between applicants. It is therefore hard to locate their abilities and fit to the company – related observations have little diagnostic value. To change the impression of an applicant, one needs unique observations that are diagnostic in that they discriminate between the focal applicant and rival applicants. For example, we might learn in a job interview that a candidate used to exhibit dishonest behavior in the course of previous employment as opposed to observing that a candidate used to be moderately friendly towards their coworkers. Presenting oneself as friendly does not really discriminate between applicants – almost every applicant will talk about their strengths, no matter how competent they are. In contrast, dishonest behavior is diagnostic of dishonest people. Both dishonest and honest people might tell the truth most of the time, but only dishonest people tell lies (Gidron et al., 1993; Reeder & Brewer, 1979; Skowronski & Carlston, 1987). Telling lies is diagnostic as it discriminates between honest and dishonest people. In the context of impression formation, this asymmetry is apparent in negative behavior being more diagnostic than positive behavior.

Additional aspects of differential interpretation and weighting of stimulus content, like expectedness, may contribute to this asymmetry between positive and negative valence. Common behaviors (characterized by high consensus rates in attributional terms; Kelley, 1967) are given less weight in impression formation than surprising and exceptional behaviors. This asymmetry applies especially to negative and extreme behaviors versus positive and moderate behaviors. Negative and extreme information is less expected in everyday contexts compared to positive and moderate information (Fiske, 1980; Fiske et al., 2007). Different positive traits or behaviors are also more redundant (i.e. highly similar to one another) than negative information items, which are more distinct and unique (Koch et al., 2016; Unkelbach et al., 2008). The asymmetry in weighting positive and negative information is manifested in several interrelated findings: Negative information is more diagnostic, less expected and more distant compared to positive information.

In addition to diagnosticity and expectedness conceived as systematic properties of the sampled stimuli, differential weighting and interpretation of stimulus content is also dependent on individuals' internal states of mind and their personal experience with the task environment. We must therefore consider the joint impact of both Brunswikian and Thurstonian sampling on diagnosticity. The reason why yoked controls and original samplers may arrive at divergent impression judgments is that the same stimulus information whose apparent diagnosticity led one person to truncate a sample may appear less diagnostic to another, yoked control person.

Regression by Thurstonian Oscillation

Fluctuations in subjectively perceived diagnosticity may originate in a variety of quasi-random oscillations of internal states of mind: variation of attentional focus, current memory context through priming, individual autobiographic associations, prior expectancies, and so forth. Because of all these dynamic activities within the judge's mind, Thurstonian oscillations can have a profound influence on impression formation, as demonstrated by Prager et al. (2018).

When sample size is determined externally and judges are not free to determine their own information search, Thurstonian sampling is simply personal and situational noise that renders the resulting impression judgments less systematic and less reliable, without exerting a measurable influence on the judgment. When it comes to self-truncated sampling, however, the impact of Thurstonian sampling increases radically. Complementing the (Brunswikian) input of sampled traits, Thurstonian oscillations modulate the subjective diagnosticity in the mind of the beholder and thereby exert a strong impact on the truncation process. Agents truncate sampling when they feel ready to make a judgment: This readiness to judge depends on both the Brunswikian stimulus input and Thurstonian variation within the individual's mind.

A suitable method to measure these dependencies on and differences in Thurstonian oscillation is the *yoked controls* design (Prager et al., 2018). A first participant in a yoked pair engages in an impression formation task based on sequential sampling. This participant determines the sample size, truncating the sample at the very moment when they feel ready to judge the target. A second, yoked participant is then presented exactly the same sample. However, this second participant (yoked control) cannot be expected to be ready for a judgment at the very same instance at which the sample was truncated by the original participant. Perfectly synchronized cognitive activities within both yoked judges are extremely unlikely. Small samples are typically truncated when the original observer experiences synergy between the Brunswikian stimulus sample and the Thurstonian states of mind. Therefore, the individual's judgment based on early truncated samples tends to exaggerate the mere Brunswikian (ecological) sample contents. The amplification of Brunswikian input through alignment with Thurstonian sampling, facilitated by the variation of small samples, is weakened for the yoked control, who cannot be expected to be exactly in the same Thurstonian state of mind at the moment of truncation. Therefore, judgments of the trend extracted from the sample should be diminished when small samples are passed on to yoked controls, compared to judgments from self-truncating partners based on their own samples. Conversely, larger samples grow so large precisely because the samples' evidence remained ambivalent to the original judge for so long due to a non-fit between Thurstonian and Brunswikian aspects of sampling. Such a non-compelling sample is likely to appear less ambivalent to the yoked judge. This diverging pattern of judgments in a yoked controls design can be conceived as regression toward the mean. Yoked controls' judgments can be expected to be more regressive (i.e., less pronounced for small samples and less conflict-prone for large samples) than the original samplers' judgments.

To set Thurstonian sampling influences apart from Brunswikian aspects of the sampling process, we use a straightforward averaging rule (Anderson, 1965), assessing the Brunswikian stimulus input as the average pretested scale value of all sampled traits. This benchmark measure can be conceived as actuarial judgment by an idealized judge (Dawes et al., 1989), who takes the stimulus input as it is and who aggregates and processes invariantly.

Prior Evidence

Preliminary evidence for the interplay of Brunswikian and Thurstonian sampling and for the moderating influence of diagnosticity was already obtained in three experiments reported by Prager et al. (2018). Participants were asked to judge target persons described by samples of traits on likeability. The relation between sample size and judgment strength (i.e. the magnitude of the likeability judgment pointing in the direction of the dominant valence) switched from positive (i.e., stronger judgments with increasing n) when sample size was pre-determined experimentally to clearly negative (i.e., stronger with decreasing n) when participants themselves could truncate their samples. Moreover, judgment strength and sample size were also negatively related in yoked controls, who saw exactly the same samples as their counterpart in the self-truncation condition, reflecting a regular Brunswikian sampling effect. However, yoked controls' judgments showed regressive shrinkage, reflecting the asynchrony of Thurstonian sampling between the self-truncating judge and the yoked control judge. Impressions informed by small samples were not as strong and impressions informed by large samples were not as weak in yoked controls, who were exposed to samples truncated by the yoked partner.

These sample size effects depended heavily on distinct indices of diagnosticity. Earlier truncation of samples, resulting in more polarized impression judgments, was facilitated by negative rather than by positive stimulus traits and by extreme rather than by moderate traits. Moderation by the "big two" was also demonstrated, meaning that enhanced diagnosticity of morality-related negative traits and ability-related positive traits (compared to morality-related positive and ability-related negative traits) led to more pronounced impression judgments (Reeder & Brewer, 1979; Skowronski & Carlston, 1987).

Present Investigation: Clarifying the Interplay of Brunswikian and Thurstonian Sampling

Despite the encouraging evidence obtained in the Prager et al. (2018) research, the conceptual distinction between Brunswikian and Thurstonian sampling and the negative relation between sample size and impression strength in self-truncated sampling remain novel and insufficiently explored. Little is known about underlying mechanisms and about functional boundary conditions and limitations of the delineated phenomena. Hence, more systematic research and theorizing are clearly needed. One might wonder to what extent self-truncated sampling involves the strategic "exploitation of good luck" (Edwards, 1965), that is, how judges exploit the variation in diagnosticity and consistency that can be expected by chance in small random samples. Or, one may speculate about the degree of asynchrony that can be expected in yoked partners, following statistical sampling theory.

One intriguing question is whether the phenomena related to Thurstonian sampling not only apply to variation between individual judges (in a yoked control design) but also to variation within the same individual across time and occasions. Thus, when original judges are later exposed to "their own" sample that they have themselves truncated on an earlier occasion, they cannot be expected to be in the same critical state of mind that had prepared them to truncate early and to form a very strong impression judgment immediately after the truncation decision. However, in this case, when participants become their own yoked controls, as it were, the regressive shrinkage of impression judgments should be less pronounced than when other individuals serve as yoked controls. In other words, when Thurstonian processes can only vary over time within participants, their influence should be weaker than when allowed to vary between individuals and over time.

Such theoretical issues demand for systematic elaboration and careful consolidation through well-designed experimental research. The present investigation aims to illuminate the interplay of Brunswikian and Thurstonian sampling by drawing on the impression-formation paradigm developed by Prager et al. (2018). A primary goal, and premise for all further ideas, is to replicate and substantiate the robustness of the sophisticated pattern of previous findings, using novel samples of participants and stimulus materials. Second, prior to conducting new experiments, we make a deliberate attempt to elucidate the theoretical implications of our sampling approach in comparative computer simulations of impression judgments based on externally determined and self-truncated trait sampling. Our simulation model also provides suitable operational definitions of our theoretical key concepts, Brunswikian and Thurstonian sampling. Third, we conducted two new experiments, providing convergent evidence for self-truncation effects obtained with different stimulus materials. Going beyond previous findings (Prager et al., 2018), we not only expose yoked control participants to samples truncated by other participants. We also create an intrapersonal version of a yoked control design, exposing participants to trait samples they themselves had truncated at an earlier stage. In this way, we decompose Thurstonian sampling into different sources of variation, namely, inter-individual variation between different participants and intertemporal variation within the same participants.

Summary of Predictions

The most basic prediction regarding Brunswikian sampling is that participants will be sensitive (a) to the true parameters of the populations from which the stimulus samples are drawn and (b) to the unique deviation of samples from their population parameters (i.e., sampling error). Judgments are therefore expected to reflect both: (a) systematic and (b) unique properties of respective samples.

Since clear-cut, converging, and extreme information is more likely to occur in small rather than large samples, and since adding more information is likely to increase ambivalence, small samples are expected to give rise to stronger impressions than larger samples. Such a negative relation between sample size and judgment strength should result from self-truncated sampling.

Our discussion of Thurstonian sampling implies that impressions triggered in yoked controls by the very same truncated samples will exhibit regressive shrinkage due to asynchrony in Thurstonian oscillation. The strength of regression depends, in turn, on the degree of divergence or asynchrony that can be expected between original judges and yoked controls. If the same person judges their own samples again in a later situation, regression can be expected to be weaker than when another person provides a judgment based on the same sample.

All influences discussed so far depend on the diagnosticity of the sampled contents. Highly diagnostic samples (negative and extreme vs. positive and moderate traits) will be truncated earlier and judged more strongly. In addition to valence and extremity, another property related to expectedness is scarcity or infrequency of occurrence. Common traits with a high occurrence rate should have less impact on impressions and truncation decisions, compared to rare traits; samples of frequent traits will be truncated later and the resulting judgments will be weaker.

Diagnosticity effects will be not only visible in pronounced impression judgments but will also carry over to higher-order phenomena: The negative relation between sample size and judgment strength depends on the presence of truly compelling information; the more diagnostic the sampled contents, the stronger is the negative correlation. Similarly, the yoked controls' judgments should exhibit more regressive shrinkage when highly diagnostic stimulus contents amplify the self-truncation effect and render the original sampler's judgment extreme.

Simulation Study

To illustrate and elucidate the emergence of a negative relation between sample size and judgment strength through self-truncation and the regressive shrinkage of this relation from self-truncation to yoked controls, we report a simulation study. The purpose of this simulation is to demonstrate the emergence of the postulated self-truncation patterns in an idealized and simplified context. Our simulation is not meant to identify a distinct cognitive mechanism applied by participants of the subsequent impression formation experiments. Yet, it certainly clarifies the statistical constraints imposed on the formation of impressions from self-truncated sampling in an exemplary fashion. In fact, there are plenty of ways to track or model self-truncated sampling. For the purpose of this simulation study, we decided on a formalization of the impression formation process by means of Bayesian inference (Edwards, 1965) from dichotomous information samples.

Although our experiments acknowledge that each piece of information might be located in a specific position over the full range of likeability between "extremely unlikeable" and "extremely likeable", we simplify sampled stimuli as (stochastically independent) dichotomous information to render the simulation more traceable and convenient. Every simulated piece of information can either be "likeable" or "not likeable". We simulate impression formation as an inference of the probability of observing the attribute "likeable", which can be accomplished within the framework of Bayesian updating of beta-distributed priors.

As long as nothing is known about the target person before sampling, the principle of insufficient reason (Savage, 1954) implies that each level of $p(\text{"}likeable\text{"})$ must be assumed to be equally likely (see flat curve of the left-most graph in Figure 1)[1].The

[1] In our experimental design participants should not deviate considerably from uniform prior beliefs, since

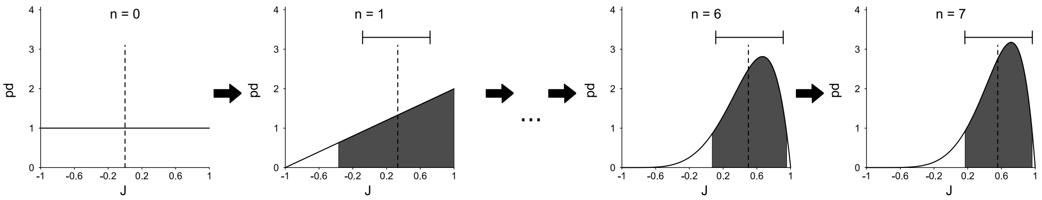


Figure 1. Illustration of the highest density interval constituting a stopping rule for the simulation of impression judgments. The algorithm is provided with a series of "likeable" and "not likeable" stimuli. The solid curve of each graph represents the posterior probability density function (i.e. an instantiation of the beta distribution) for increasing sample size from $n = 0$ to $n = 7$ (left to right). The vertical dashed line represents the (posterior) mean of likeability reflected in the current sample. The 90% posterior highest probability density interval (grey area) is compared against the threshold value t (width of the line segment above the distribution curves; here $t = .4$), which is reached at $n = 7$ in this example (far right plot). All prior and posterior values (which express a p(”*likeable*”) value) are transposed to a [-1, 1] interval, namely into impression judgment J.

mean of this prior belief about the target likeability is 0, indicating a neutral or indecisive impression. This likeability belief is then updated after each piece of information encountered in a sequential sample of traits (i.e. "likeable" vs. "not likeable"). For instance, if the first trait is positive, the posterior probability *p("likeable"|Trait 1)* is now more likely high than low, as evident from the negative skew of the density function in the second chart of Figure 1. Further steps of updating from an exemplar sample follow in Figure 1. To determine a distinct judgment from the posterior density distribution of *p("likeable"|Sampled Traits)*, we rely on the posterior mean (represented by dashed vertical lines in Figure 1).

But how will the simulated judge decide when to stop sampling? The monadic structure of the impression formation task calls for a different answer than the dyadic choice between two options (samples) assumed in most prior work on optional stopping in decision making (Kahan et al., 1967; Payne et al., 1988; Simon, 1955) and statistics (Wald, 1947). Stopping rules for choice must be sensitive to the variance between competing options: samples are typically truncated as soon as the sampled evidence clearly favors one option over others. An impression judgment task, in contrast, calls for a different stopping rule. A natural moment to stop sampling is when the growing evidence on the target stabilizes and settles on a coherent impression. A Bayesian updating framework offers such a stopping criterion in terms of the subjective certainty of a growing impression, which can be well expressed by the shape of an updated density curve.

Returning to Figure 1, we can see that the interval shaded in dark grey, which marks the interval of the 90% highest density (i.e. the narrowest 90% interval possible) becomes narrower with increasing sample size n, from the left to the right chart. Sampling is truncated when this highest-density interval shrinks to less than a critical width limit t (indicated by the horizontal lines above the density curves) that is, when posterior p(”*likeable*”) converges on a reasonably compact high-density range.

Updating the beta distribution provides a simple and straightforward method to formally simulate the stopping rule and the resulting posterior mean of the depicted algorithm. The beta distribution requires two shape parameters and β to specify the probability density function of p(”*likeable*”). Assuming full ignorance at the beginning, both shape parameters are initially set to $\alpha_0 = \beta_0 = 1$, resulting in a uniform prior

populations with different expected values of valence (i.e. extremely negative – moderately negative – moderately positive – extremely positive; see methods section for details) from which samples are drawn are randomly exchanged after each consecutive trial.

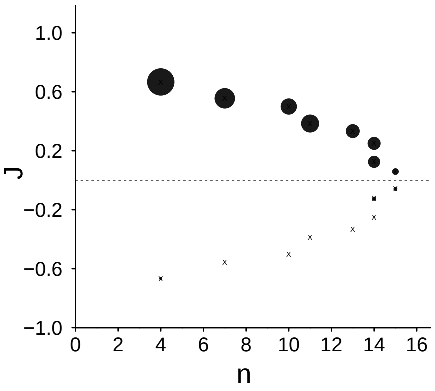


Figure 2. Simulation results of impressions generated by the Bayesian sequential estimation procedure applying a probability-density-dependent truncation rule. The graph plots impression strength J against sample size n. Samples are generated by independent sequential random draws of "likeable" vs. "not likeable" with the probability of the attribute "likeable" of $p = .75$. The algorithm starts from uniform priors $Beta(1, 1)$. $N = 200,000$ sampling trials were simulated. Sampling was truncated whenever the range of the 90% posterior highest density interval was smaller than the threshold value $t = .4$, which served as a moderate solution in the trade-off between stopping all samples extremely early and stopping only at very large sample size. Dot sizes are proportional to the frequency of occurrence. The overall correlation between sample size $ln(n)$ and impression strength J is $r = -.80$.

distribution, expressing that every possible likeability value is expected equally probable prior to seeing any evidence. After each sampled stimulus trait (i.e., evidence supporting either "likeable" or "not likeable"), the two shape parameters are updated: $\alpha = \alpha_0 + k$ and $\beta = \beta_0 + n - k$, where n is the current sample size and k the number of "likeable" traits. To extract a likeability judgment from the updated beta distribution, we transform the judgment scale (i.e., the horizontal axis in the charts of Figure 1) to a range from -1 ("extremely unlikeable") to +1 ("extremely likeable")[2]. The resulting judgment J amounts to $J = 2M - 1$ if true $p(''likeable'') \geq \frac{1}{2}$ or $J = -2M + 1$ if $p(''likeable'') < \frac{1}{2}$, with the posterior mean calculated as $M = \frac{\alpha}{\alpha+\beta}$.

Applying the discussed procedure, a first series of simulations produced a pronounced negative relation between sample size and impression strength (this relation is positive for externally determined sample size in the Bayesian framework for the given parameters). When samples were drawn from a predominantly negative population opposed to a predominantly positive population, impressions were more negative for smaller n and more moderate for larger n, so we can still see a negative relation of sample size and impression strength (i.e. deviation from neutrality in the positive or negative direction). The relation disappears when the target is perfectly neutral (i.e. $p = \frac{1}{2}$), since all impressions are then symmetrically distributed around the center. Yet, even in this case, small samples yield more extreme judgments – they just cancel each other out because they symmetrically deviate in both directions from the center. Figure 2 displays simulation results from an exemplary set of parameters.

To introduce Thurstonian sampling to the simulation procedure, we assume that both the updating of the posterior likeability distribution and the threshold parameter setting are affected by oscillations within the judging individual's mind. Because Thurstonian oscillations reflect a variety of different cognitive influences and activities, we resort to

[2] Additionally, values are reversed when the actual (population) $p(''likeable'')$ value is smaller than .5, that is a predominantly (truly) unlikeable target.

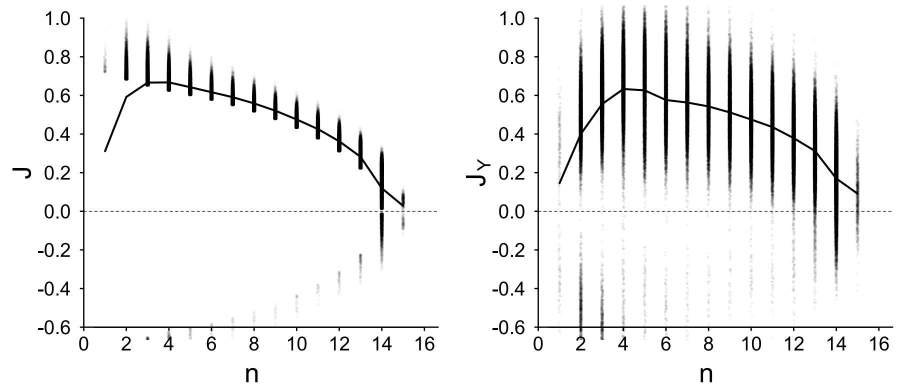


Figure 3. Simulation results for self-truncated samples (left) and corresponding yoked control samples (right): likeability judgment strength J dependent on sample size n. Basic input parameters are identical to those of Figure 2. At each updating step (n), normally distributed Thurstonian noise was applied to the beta distribution shape parameters α and β (oscillation SD = .2). For yoked controls, noise was removed and independently re-added the same way. Sampling was truncated whenever the self-truncating judge's posterior highest density interval (including noise) was narrower than t = .4. Impression strength J was calculated according to the respective mean of the posterior distribution. The clear-cut negative correlation ($r = -.59$) of impressions strength and self-truncated (log) sample size $ln(n)$ regresses to $r = -.38$ for yoked controls.

a simplifying model by adding normally distributed noise with an expected value of 0 and a certain standard deviation (here $SD = .2$)[3] in every updating step to the shape parameters (and β) of the beta distribution. More precisely, for each updating step (or sample size n), a new noise component was generated and applied to the beta shape parameters (added to and subtracted from β), resulting in fluctuation in how the new piece of informational input is currently perceived (in terms of likeability).

To simulate the regression effect that distinguishes yoked controls from original samplers, we removed the Thurstonian fluctuation from the original sampler's shape parameters and β and added a new random-fluctuation component supposed to represent a yoked control's independent Thurstonian state of mind. The results corroborate the prediction that the correlation between sample size n and judgment strength J shrinks noticeably between self-truncation (left in Figure 3) and yoked controls (right in Figure 3).

Although a variety of parameter settings converged in producing a clearly negative relation between sample size and impression strength, certain distinct cases set boundaries to the generality of the phenomenon. The effect vanishes when thresholds are set to extreme values (e.g., when all samples are either truncated very early or at extremely large sample sizes where the effect fades out). In contrast, strongly informed priors approximating the sample's true p(*"likeable"*) only shift the truncation point towards earlier truncation, but do not actually affect the described sample size effects: Reducing surprise by introducing accurately informed priors triggers early truncation, but does not change the relationship between sample size and likeability judgments. However, strong priors deviating markedly from the true p(*"likeable"*) will obscure patterns and may, sometimes, even reverse the effect. In this case, incorrect priors may cause a shift in the location of the posterior distribution in a direction not supported by the sample. For convenience we have omitted adding Thurstonian noise to the truncation threshold

[3] The oscillation SD of .2 also takes a medium position between no noticeable influence of Thurstonian elements on the sampling procedure and an entirely noisy and obscured pattern of results.

t, too. When the threshold value is also subject to random fluctuation, the sample size pattern is flattened (i.e. the correlation between sample size and strength of judgments shrinks) and samples tend to be truncated earlier, since this kind of noise causes a tendency towards smaller samples due to instances of randomness-driven early truncation. The regressive shrinkage between self-truncated and yoked control trials however is not affected by this additional source of Thurstonian noise. Details of these extensions and limiting conditions are provided in the supplemental materials.

To conclude, our simulations demonstrate and clarify how it is possible that impressions from self-truncated samples tend to overestimate existing population trends when sample size is small and tend to slightly underestimate existing trends when samples are large. These sample-size dependent biases are less pronounced in simulated yoked control judgments, where sample size does not relate as strongly to judgment strength as for the simulated self-truncated sampling. This phenomenon does not contradict common expectations that estimations become increasingly stable and certain when sample size increases, but is perfectly in line with it. Smaller samples are more likely than larger samples to produce the kind of strong deviations from population parameters that enable truncation at moments that allow samples to exaggerate underlying trends. These results support the predictions derived from our theoretical framework, they additionally refine the expected empirical patterns and effects.

Experimental Evidence

One purpose of Experiment 1 was to disentangle a confound of the Prager et al. (2018) yoked controls design: The yoked controls' regressive judgments might have reflected a notable procedural difference rather than only Thurstonian sampling differences. Whereas self-truncating participants were actively involved in sample solicitation, letting the sequential sample unfold and truncate in a self-determined way, their yoked controls passively observed an externally provided sample. Experiment 1 introduces a crucial refinement in the yoking procedure: some participants see their own samples again (the "self"-condition), whereas others observe samples truncated earlier by another (i.e. partner) participant (the "other"-condition). This allows for observing "self" and "other" yoked partners at two levels of dissociated Thurstonian processes. Presentation mode or involvement in the sampling procedure is however identical for "self" and "other" yoked participants. Both participants working on "self" and "other" yoked controls trials have previously worked on self-truncated sampling. This parallel procedure rules out possible confounds of the former research design, varying only the source of sample truncation for the yoked control sampling. Procedural differences between active sampling and truncation versus passive sample observation cannot account for differential impression judgments because both types of yoked partners ("self" and "other") observe the stimuli in the same passive and externally determined way. Still, the theoretical distinction between inter-temporal and inter-individual sources of Thurstonian variation predict a stronger sample-size dependence for judgments of "self"- compared to "other"-yoked controls. Between the self-truncated judgments and the "self"-yoked controls only time changes, whereas for the comparison between self-truncated judgments of the "self" condition and "other" yoked controls, both time and individual change.

Experiment 1

Methods

General Task and Stimulus Materials. Participants were presented with sequential samples of trait adjectives. They were informed that the traits had been reported by members of a seminar characterizing their fellow students. Based on each sample of traits related to a single target person, participants rated the likeability of the target.

Trait samples were drawn from a pool of fifty-seven trait adjectives used by Prager et al. (2018), originally taken from the Berlin Affective Word List – Reloaded (Võ et al., 2009). From the entire pool, four potentially overlapping population sets containing 30 traits each were drawn[4]. Two sets of predominantly negative traits (proportion of positive traits of .20 and .33) were mirrored by two sets of mainly positive traits (proportion positive of .80 and .67). Within each of these two trait sets of opposite valence, one set represented moderate valence and one extreme valence. Mean valence scale values were held approximately symmetrical, variance and skewness approximately equal across both factors, positive versus negative and moderate versus extreme valence (see Appendix).

Participants worked on nine trait samples from each of the four stimulus sets in both blocks of the experiment, summing up to 36 samples per block shuffled in random order, thus yielding an orthogonal within-participant manipulation of positive versus negative and moderate versus extreme targets. During sampling, trait words were displayed sequentially in black letters (font size 20pt) on white background, filling the screen (from top left to bottom right) up to 16 word slots of an invisible 8 x 2 grid. All traits of a sample remained onscreen, but only the most recent trait appeared in full black while all previously presented traits were reduced to grey after the onset of the next trait.

Procedure. The experiment was the first in a series of three studies included in a one-hour lab session at Heidelberg University. It consisted of two blocks separated by an unrelated experiment on causal impact ratings. The entire experiment was controlled by a Java-program.

Participants were assigned to one of two experimental conditions, "self" or "other". The first participant assigned to a computer workplace in the lab was assigned to the "self"- condition. The second participant assigned to the same computer then served as the yoked control participant. The following participant was assigned to the "self"-condition again and so forth. All participants, regardless of their experimental condition, completed two blocks. In the first block, they received samples of traits. At each sampling step, they themselves could actively control whether to consider another trait (by pressing the space bar) or to truncate sampling at the current state (by pressing the enter-key). Very brief summaries of instructions remained visible throughout sampling in small font on bottom of the screen. During this first block, both "self" and "other"-participants could gather samples of up to 16 traits drawn randomly from the same ordered set of 16 stimuli. Participants of both conditions were instructed to gather traits until they felt ready to make a judgment. Since all samples were self-truncated, sample size differed depending on individual truncation decisions between "self" and "other"-participants, despite the yoked sample content. Participants worked on a total of 36 samples during the first block.

After truncation of a sample by pressing the enter-key, a continuous rating scale (18 cm long) appeared below the listed traits, ranging from "highly unlikeable" to "highly likeable". Immediately afterwards, a rating of confidence in the preceding likeability

[4] These population sets were not identical to those used by Prager et al. (2018).

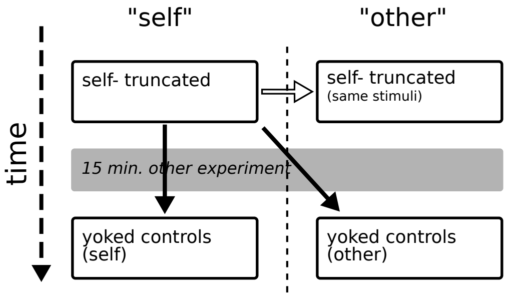


Figure 4. Schematic illustration of the design used for Experiment 1.

judgment was requested, using a similar graphical scale underneath the likeability scale, along with the labels "very unsure", "neutral", and "very sure" (from left to right). Ratings were submitted by simply clicking on the scales. Judgments were visualized by a tick (small vertical bar) on the scale; its position could be corrected. After both judgments were registered and participants confirmed their entries by clicking a button – the screen cleared and the next trait sample started.

Ten to 15 minutes later (after an experiment involving causal-impact ratings), participants of both conditions worked on a second block of trait sampling and impression judgments. This time, they were not free to decide for themselves whether to extend the sample or truncate; rather they passively observed the trait samples that had been recorded during the first stage of the "self"-condition. The visual presentation mode of the sample remained identical to the previous block. Samples grew at a fixed pace of one second per trait, but stopped automatically. When the pre-determined end of a sample was reached, participants provided their judgments in the same way as during the first block. Both conditions received the 36 trait samples that the "self"-condition had truncated during the first block in a new random order. The study design is illustrated by Figure 4.

Participants. Ninety-six participants (71 female) were recruited from the online pool "Studienportal" at Heidelberg University. Eighty-nine participants were students; their age ranged from 17 to 63 years (Mean = 24.99 years). Four participants were excluded; two had always chosen the same sample size and two had heavily extended median response times (more than four standard deviations above the average individual median). Their yoked counterparts also had to be excluded. The remaining $N = 88$ participants were included in the analysis, of which $N = 42$ pairs could be formed; four participants of the "self" condition did not have a yoked control partner because they were not succeeded by another participant.

Results

Systematic Nature of Impression Judgments and Sensitivity to Sampling Input. Before we turn to the central hypothesis tests, it is necessary to assess the effectiveness of the manipulation and the sensitivity of participants' likeability judgments to the sampling input, according to the Võ et al. (2009) valence norms. We used the average valence norms of stimulus traits (Anderson, 1965) as a benchmark for comparing judgments to the Brunswikian sample input.

Table 1. Regression of Likeability Judgments on Population Set Average and Sampling Error

Predictor	$M\beta$ (SD)	Consensus	t	df	p-value
Block 1: self					
Population average	.61 (.15)	100%	27.38	44	$< .001$
Sampling error	.38 (.15)	98%	17.65	44	$< .001$
Block 1: other					
Population average	.65 (.16)	100%	25.83	42	$< .001$
Sampling error	.39 (.13)	100%	19.17	42	$< .001$
Block 2: self					
Population average	.62 (.20)	98%	21.02	44	$< .001$
Sampling error	.37 (.14)	100%	17.93	44	$< .001$
Block 2: other					
Population average	.63 (.18)	100%	23.27	42	$< .001$
Sampling error	.25 (.13)	100%	12.63	42	$< .001$

Note. Consensus refers to percentage of participants whose β matches the sign of the general trend. As both predictors are orthogonal, regression weights β (approximately) equal zero-order correlations r.

As each trait sample was drawn from one of four population sets, each individual judgment was subject to the systematic influence of average valence of that population set. In addition to this experimentally controlled systematic property, however, each judgment also depended on *sampling error*, that is, the Brunswikian sample's deviation from the population mean.

Two predictors, the systematically manipulated population set average and the unsystematic sampling error, both calculated from trait valence norms were included in a hierarchical multiple regression, using participants' likeability judgments as criterion. The resulting regression weights in Table 1 indicate that both predictors accounted for a substantial portion of systematic judgment variance, testifying to the high data quality and the judges' generally high degree of sensitivity to the experienced trait input. Mean individual proportions of explained variance are $R^2 = .60$ (block 1, "self"), $R^2 = .64$ (block 1, "other"), $R^2 = .62$ (block 2, "self"), and $R^2 = .52$ (block 2, "other").

Relation of Sample Size and Judgment Strength. For a test of the central hypothesis, we calculated the hierarchical regression of judgment strength J on sample size n. Judgment strength J is equal to the likeability judgment for positive population sets; for negative sets, the sign is reversed such that positive (negative) J scores always indicate the strength of judgments pointing in the correct (incorrect) direction. For all analyses of the relation between n and J, sample size n is rescaled by the natural logarithm to $ln(n)$, which is better suited for testing the expected results pattern. All results and statistical conclusions remain unaffected by this logarithmic transformation. Table 2 summarizes the corresponding statistics; Figure 5 illustrates the same data graphically.

Data from all blocks converge in demonstrating a stable and highly consensual negative relation between $ln(n)$ and J. This replicates the results of Prager et al. (2018): Impression judgments reflected the prevailing valence more strongly when sample size was small rather than large.

Regressive shrinkage is also apparent from these results: Individual correlation coefficients of the relation between $ln(n)$ and J showed a tentative shrinkage from $r_{ln(n),J} = -.30$ to -.27, ($SD = .11$, $t(44) = 2.08$, $p = .044$, 64% consensus) from the first to the second block within the "self" condition. This reflects regressive shrinkage of

Table 2. Mean Correlations (and Standard Deviations) Between Individual Participants' Judgment Strength J and the Natural Logarithm of Sample Size $ln(n)$ Across Trait Samples

Experimental condition & block	$M\beta$ (SD)	Consensus	t	df	p-value
Block 1: self	-.30 (.22)	89%	-9.24	44	< .001
Block 1: other	-.28 (.21)	93%	-8.75	41	< .001
Block 2: self	-.27 (.24)	84%	-7.37	44	< .001
Block 2: other	-.17 (.20)	81%	-5.50	42	< .001

Note. Blocks 1 and 2 of the other condition differ in degrees of freedom because one participant of Block 1 had to be excluded due to invariant sample size.

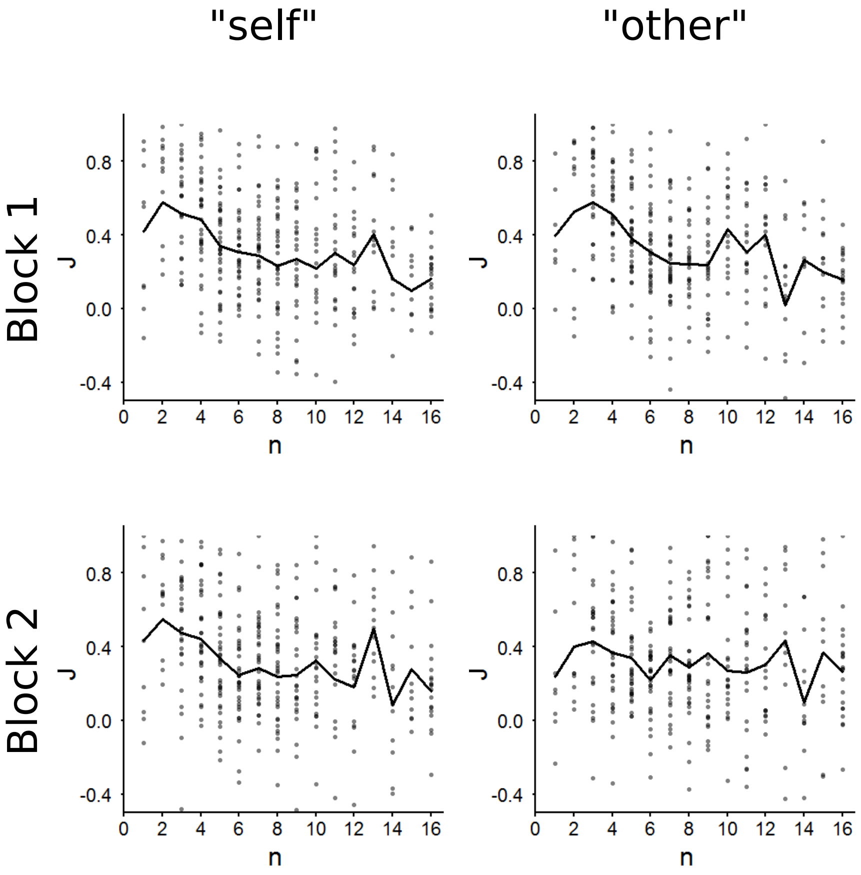


Figure 5. Judgment strength J plotted against size of self-truncated samples n of the "self" condition and yoked control samples for the "self" and "other" condition. Dots represent individually averaged judgment strength values per sample size, lines cross averaged individual means per sample size.

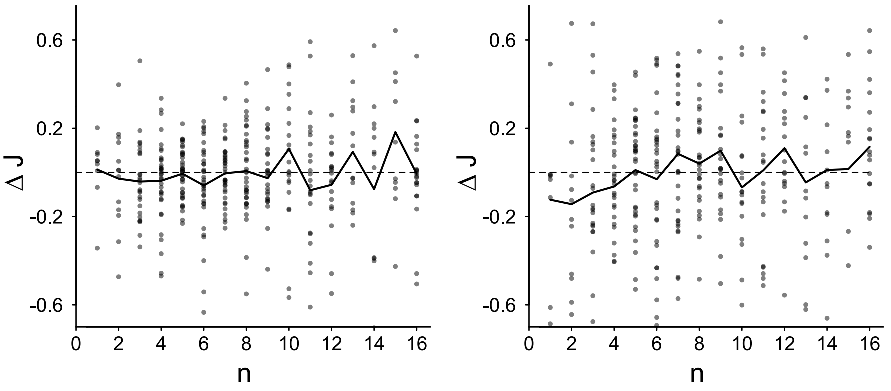


Figure 6. Differences in judgment strength ΔJ between the first and second block of the "self"-condition (left hand side) and between conditions "self" and "other"-yoked controls (i.e. between conditions of the second block; right hand side). Dots represent individual mean values per sample size, solid lines the averaged individual means per sample size.

judgments based on the judges' own formerly truncated trait samples. The negative correlation decreased more strongly from $r_{ln(n),J} = -.27$ to -.17 (SD = .25, $t(41) = 2.51$, $p = .016$, 64% consensus) when comparing "self" and "other" yoked controls' judgments during the second block.

In the twofold yoked design, participants' judgments can be organized in triplets. We could analyze two quantities of regressive shrinkage by forming two types of difference scores: between judgments of self-truncated samples and yoked controls samples within the "self"-condition, and between judgments of "self" and "other"-yoked controls. Figure 6 reflects the expected regressive patterns: When, in the second block, judges were exposed to exact copies of self-truncated samples from the "self"-condition's first block, the originally stronger judgments from small samples shrunk to somewhat weaker judgments, whereas the originally weaker judgments from larger samples tended to increase somewhat. The same regressive pattern is evident in slightly positive correlations between (log) sample size $ln(n)$ and the change in judgment strength ΔJ. Both, the correlation for the judgment changes in the "self" condition (mean $r_{ln(n),\Delta J}$ = .06, SD = .17, $t(44) = 2.32$, p = .025, 64% consensus) and the correlation for the difference between "self" and "other" yoked controls (mean $r_{ln(n),\Delta J}$ = .09, SD = .26, $t(41) = 2.17$, $p = .036$, 62% consensus) tended to be positive – a reduction of originally negative correlations.

Diagnosticity. Our theoretical conception emphasizes diagnosticity as a chief determinant of Thurstonian sampling effects. For a test of the influence of trait diagnosticity on truncation and resulting impression judgments, we conducted two hierarchical regression analyses, using self-truncated n and judgment strength J as criteria, and three measures of diagnosticity as predictors: population set valence (expected diagnosticity difference: negative > positive), population set extremity (extreme > moderate), and trait word frequency (rare > common) according to norms (frequency count per one Million words of written language; Võ et al, 2009).

The regression of n revealed that participants truncated their samples earlier for both negative and extreme population sets compared to positive and moderate sets (Table 3 and 4). From the zero-order correlations it is evident that (low) average word frequency is also a predictor of early truncation but this relation is mostly absorbed by population

Table 3. Regression of Self-Truncated Sample Size $ln(n)$ on Indicators of Diagnosticity: Population Set Valence, Extremity, and Expectedness Indicator Sample Mean Word Frequency

Predictor	M r (SD)	Consensus	M β (SD)	Consensus	t	df	p-value
Block 1: self							
Population set extremity	-.11 (.19)	80%	-.11 (.20)	76%	-3.74	44	<.001
Population set valence	.25 (.19)	91%	.23 (.23)	84%	6.72	44	<.001
Sample mean word frequency	.18 (.18)	84%	.04 (.20)	58%	1.32	44	.194
Block 1: other							
Population set extremity	-.06 (.17)	62%	-.07 (.18)	67%	-2.42	41	.020
Population set valence	.18 (.23)	79%	.15 (.28)	64%	3.48	41	.001
Sample mean word frequency	.14 (.20)	81%	.05 (.25)	62%	1.37	41	.178

Note. Second block data were perfectly redundant with Block 1 self results.

Table 4. Diagnosticity/Expectedness Predictor Inter-Correlations

Predictor	M r (SD)	Consensus
Population set extremity and valence	0 (0)	–
Population set extremity and sample mean word frequency	-.03 (.12)	60%
Population set valence and sample mean word frequency	.62 (.08)	100%

Note. Analyses performed on data from the self-condition's first block. Predictor intercorrelations for the other conditions are virtually identical.

set valence, which constitutes a highly correlated but stronger predictor.

The other regression analyses showed that judgment strength J increased in both blocks of both conditions when underlying population sets were extreme and negative rather than moderate and positive. Average sample word frequency also predicted J; samples containing infrequent traits solicited stronger judgments (in the correct direction) than samples of frequent traits. However, as in the previous analysis, the predictive value of the word-frequency predictor vanished when population set parameters were included in the multiple regression.

We hypothesized not only an impact of diagnosticity on sample size n and judgment strength J but also on the relation between these two variables $r_{ln(n),J}$. The pertinent results appear in Table 5 and Figure 7. Whereas (negative) valence is highly predictive of individual judges' (negative) $r_{ln(n),J}$, across conditions and blocks, the extremity index contributes little to explaining the impact of diagnosticity on the accentuation of judgments informed by small samples.

Judgment Confidence. Correlations $r_{ln(n),c}$ between sample size $ln(n)$ and subjective confidence c corroborate the impact of diagnosticity. Consistently negative correlations indicate that smaller trait samples informed more confident impression judgments than larger samples.

The relation between $ln(n)$ and c was subject to the same regressive shrinkage from self-truncated samples to yoked controls as the relation between $ln(n)$ and J. Between the two blocks of the "self" condition (self-truncated, then yoked controls), the $r_{ln(n),c}$ correlations did not shrink noticeably (mean $\Delta r_{ln(n),c} = .03$, $SD = .24$, $t(44) = .72$, $p = .475$, 51% consensus). Comparing yoked control participants in the "self" versus "other" condition, however, revealed substantial shrinkage (mean $\Delta r_{ln(n),c} = .15$, $SD = .37$, $t(41) = 2.74$, $p = .009$, 67% consensus).

Judgment confidence was also subject to the same impact of diagnosticity as judgment strength. As is evident from Table 8, mean c was higher for samples drawn from extreme and negative population sets consisting of infrequent traits than for samples from moderate and positive population sets consisting of frequent traits. Again, the

Table 5. Regression of Judgment Strength J on Indicators of Diagnosticity: Population Set Valence, Extremity, and Expectedness Indicator Sample Mean Word Frequency

Predictor	M r (SD)	Consensus	M β (SD)	Consensus	t	df	p-value
Block 1: self							
Population set extremity	.16 (.14)	87%	.17 (.15)	87%	7.71	44	$< .001$
Population set valence	-.37 (.36)	82%	-.41 (.41)	82%	-6.70	44	$< .001$
Sample mean word frequency	-.19 (.24)	78%	.08 (.21)	71%	2.50	44	.016
Block 1: other							
Population set extremity	.20 (.14)	93%	.20 (.14)	93%	9.65	42	$< .001$
Population set valence	-.35 (.34)	86%	-.38 (.36)	88%	-6.87	42	$< .001$
Sample mean word frequency	-.20 (.26)	84%	.05 (.19)	58%	1.69	42	.098
Block 2: self							
Population set extremity	.17 (.13)	93%	.17 (.13)	89%	8.58	44	$< .001$
Population set valence	-.43 (.37)	87%	-.45 (.40)	82%	-7.59	44	$< .001$
Sample mean word frequency	-.24 (.27)	84%	.04 (.22)	58%	1.33	44	.190
Block 2: other							
Population set extremity	.20 (.15)	88%	.20 (.15)	88%	8.88	42	$< .001$
Population set valence	-.32 (.32)	81%	-.35 (.38)	84%	-6.10	42	$< .001$
Sample mean word frequency	-.17 (..22)	81%	.06 (.20)	65%	1.87	42	.069

Table 6. Regression of $r_{ln(n),J}$ on Population Set Valence and Extremity

Predictor	β (SE)	Consensus	t	df	p-value
Block 1: self					
Population set extremity	-.08	57%	-1.05	170	.294
Population set valence	.30	68%	4.13	170	$< .001$
Block 1: other					
Population set extremity	-.11	62%	-1.54	164	.132
Population set valence	.34	69%	4.59	164	$< .001$
Block 2: self					
Population set extremity	-.02	50%	-.25	170	.803
Population set valence	.23	61%	3.13	170	.002
Block 2: other					
Population set extremity	-.00	48%	-.01	164	.99
Population set valence	.37	79%	5.11	164	$< .001$

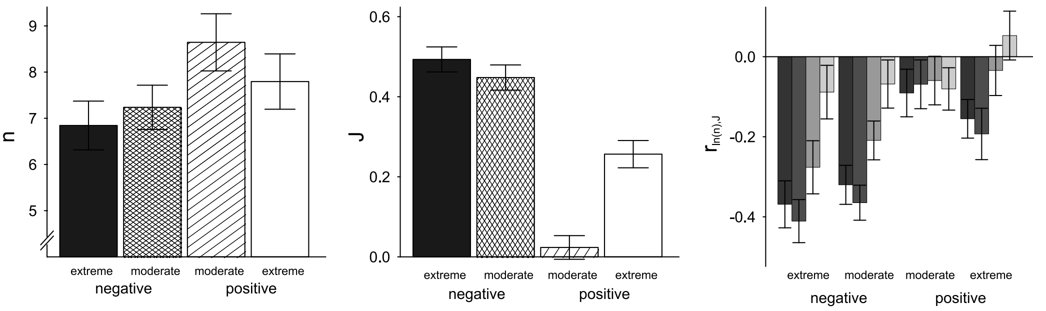


Figure 7. Mean sample size n, mean judgment strength J and mean correlation $r_{ln(n),J}$ between n and J (left to right) as a function of extremity (moderate vs. extreme) and valence (negative vs. positive). Error bars indicate standard errors of individual means. The bar chart for $r_{ln(n),J}$ is broken down by sampling conditions "self", "other", "self-yoked", and "other yoked" (left to right and dark to light grey shading).

Table 7. Mean Correlations $r_{ln(n),c}$ Between the Natural Logarithm of Sample Size $ln(n)$ and Self-Reported Confidence in Impression Judgments c

Experimental condition & block	M $r_{ln(n),c}$ (SD)	Consensus	t	*df*	*p*-value
Block 1: self	-.25 (.31)	76%	-5.50	44	< .001
Block 1: other	-.29 (.28)	83%	-6.83	41	< .001
Block 2: self	-.23 (.24)	78%	-6.33	44	< .001
Block 2: other	-.06 (.25)	64%	-1.56	41	.127

Table 8. Regression of Confidence in the Judgment c on Indicators of Diagnosticity: Population Set Valence, Extremity, and Expectedness Indicator Sample Mean Word Frequency

Predictor	M r (SD)	Consensus	M β (SD)	Consensus	t	*df*	*p*-value
Block 1: self							
Population set extremity	.15 (.17)	78%	.15 (.17)	80%	5.96	44	< .001
Population set valence	-.16 (.22)	76%	-.18 (.26)	69%	-4.57	44	< .001
Sample mean word frequency	-.10 (.18)	73%	.02 (.21)	53%	.59	44	.559

Note. Only results from the first block of “self” condition are listed. Results from all other blocks are highly convergent.

predictive value of the word frequency measure was absorbed by the other diagnosticity indices in the multiple regression (see discrepancy between zero-order correlations and regression coefficients).

Correlations between sample size and confidence were not (noticeably) influenced by diagnosticity (Table 9). Yet, the overall pattern converged with that of judgement strength.

Discussion

The entire pattern of findings testifies to the sensitivity of impression judgments to systematic and unsystematic variation in the sampled information. Judgments were not only highly sensitive to the valence of the population set from which the trait samples were drawn, but also to the sampling error or deviations of the sampled traits from the population mean. The same regular influence of diagnosticity on impression judgments was manifested in judgment strength J, the subjective confidence c of impression judgments, and the sample size n at which participants ended sampling because they deemed it sufficient. High trait diagnosticity not only led to early truncation at small n and to sensitive judgments J at high levels of confidence c but – consistent with the sign of these separate influences – also to distinct negative relations $r_{ln(n),J}$ and $r_{ln(n),c}$ between n and judgment strength and confidence, along with positive relations between J and c. The entire pattern determined by the sampling and diagnosticity of stimulus traits was observed at a high rate of consensus across most individual participants, providing convergent validation for the findings obtained previously (Prager et al., 2018).

However, crucially, these highly sensitive impression judgments were not exclusively determined by the meaning and diagnosticity of the traits inherent in Brunswikian stimulus samples. They also depended regularly on Thurstonian sampling processes

Table 9. Regression of $r_{ln(n),c}$ on Population Set Valence and Extremity

Predictor	M β (SD)	Consensus	t	*df*	*p*-value
Population set extremity	.02	45%	.25	170	.804
Population set valence	.09	64%	1.12	170	.264

within the judges' mind. This less obvious, Thurstonian, influence is evident from the regressive shrinkage in the yoked control participants' judgments J and their negative correlations $r_{ln(n),J}$, relative to self-truncating participants, whose impression judgments were based on exactly the same trait samples. The only reasonable explanation for regressive shrinkage is that yoked controls were not as ready or well-prepared to make a judgment at the time of truncation as the self-truncating judges, whose internal sampling oscillations were obviously aligned with the instance of sample truncation.

Further support and more refined evidence for the interplay of Brunswikian and Thurstonian sampling comes from the newly introduced manipulation of "self"-yoked versus "other"-yoked controls. Re-presenting participants with their own samples truncated in a previous block also resulted in regressive judgments and less pronounced negative relations $r_{ln(n),J}$ between n and J. However, the regression effect after self-yoking was smaller than the regression observed in the other-yoking condition, when participants were exposed to the samples truncated in an earlier block by another person. In other words, the amount of variance explained by Thurstonian sampling was smaller within participants over time than between participants, when temporal and interpersonal variance came together.

The present experiment and the Prager et al. (2018) research both used the same general set of trait adjectives and the same norms. Population sets (i.e. the populations from which samples were drawn) were not identical between experiments, but they overlapped considerably. Furthermore, Experiment 1 was conducted in two distinct blocks with a clearly different task setting (self-truncated vs. yoked controls tasks). Yoked controls' judgments may have been generally less extreme (for small samples especially) compared to judgments using self-truncated samples. Thus, the distribution of judgments on an underlying reference scale might have differed between Block 1 (self-truncation) and the yoked controls of Block 2 (Thurstone, 1927), due to a generally broader range of judgments in the self-truncated sampling phase. Yoked control participants may have therefore expanded their scale usage, because they generally experienced extreme impressions less frequently compared to the first block of self-truncated sampling (Parducci & Perrett, 1971; Stewart et al., 2006).

Experiment 2

To fix these issues and to further substantiate our theoretical approach, we modified the setup for Experiment 2 in two respects: First, we generated entirely new, natural and non-redundant traits in a pilot-study. Replacing traits entirely helps to control for material artifacts and enhances the value of replicated findings. Second, we switched to a full within-participants-within-blocks design, in an attempt to deal with the issue of differential scale usage. By intermixing self-truncated and yoked samples in the same series, the reference scale experience of judgments based on both types of (self-truncated and yoked) samples is held constant.

Methods

Materials and pre-study procedures. In order to generate a new trait population, and also to fit the cover story more realistically, we asked 18 students from a seminar to write down trait words using paper and pencil. They were instructed to think of traits of someone they like and someone they dislike. Liking and disliking traits were written down in separate, successive blocks, administered in random order. Blocks of

12 lines were provided for traits of liked and disliked persons. This resulted in 297 trait adjectives, from which we excluded negations by prefix (German “un-“). We then randomly selected a total of 120 trait words as basic stimulus material for Experiment 2.

In the next step, we conducted an online pre-study optimized for running on mobile device browsers in order to achieve context-independent valence norms comparable to those used in the previous experiment (BAWL-R norms by Võ et al., 2009). Participants were asked to rate the valence of the trait generally, detached from a specific situation. Traits were presented one after another in bold face in the top center of the screen along with a 4.5 cm rating scale ranging from “very negative” to “very positive” below (without default). The study took on average less than 10 minutes to complete. Forty-four participants completed the pre-study; 28 were female; their age ranged from 18 to 75 (mean = 34.77). Forty-one participants identified as native German speakers; the remaining three rated their language skills in German as “fluent”. One participant was excluded because of too large inconsistency of their ratings with judgments of other participants: The correlation of this participant’s ratings with all other ratings was more than three standard deviations below the mean of this measure of all other participants. After this exclusion (pairwise) interrater correlations were on average r = .79 (.77 without exclusion).

In the next step, we added frequency-norms to the pool of traits. We retrieved the relative frequency per one million written words taken from the online-database “Deutsches Referenzkorpus” (Institut für Deutsche Sprache, 2017). For spoken language we consulted the database “Datenbank für gesprochenes Deutsch” (Institut für Deutsche Sprache, 2014). Resulting word frequencies from written and spoken language were highly redundant (r = .93). For data analysis, we merged them by averaging the natural logarithm of both z-standardized values.

Analogous to Experiment 1, four overlapping population sets (showing symmetric properties in predominant valence and extremity with proportions of positive traits of .20, .33, .67, .80) were formed, from which random samples were drawn throughout the experiment (see Appendix for more details).

Design and procedures. The general procedures of sampling and likeability judgments of Experiment 1 remained unchanged, except for a shift to a full repeated-measures design. Each participant now provided impression judgments based on both self-truncated samples and, on other trials, samples truncated by another (preceding) participant. Thus, the self-truncated samples of one participant were passed on, as other-yoked samples, to the subsequent participant at the same computer workplace, whose own self-truncated samples were again passed on to a subsequent participant, and so forth (see Figure 8). The first participant in each chain who could not receive samples from a preceding participant was presented with a randomly truncated sample (not included in analysis but necessary for balancing the experimental design). Thus, the new repeated-measures-design only allowed for “other”-yoked controls, in the terminology of Experiment 1, but not for “self”-yoked comparisons.

The experiment was controlled by a Java-program. Participants were welcomed and asked to provide demographic information. After extensive instructions, before the participants started sampling, the entire task was briefly summarized on one slide. The experiment was run either in the 4th position of a one-hour session that included five experiments or in the 2nd position of a one-hour session containing four experiments. Research took place in computer laboratories at Heidelberg University.

Prior to each trial, one of two icons signaled whether the following task involved self-truncation or passive observation (as yoked control; Figure 9). On self-truncation

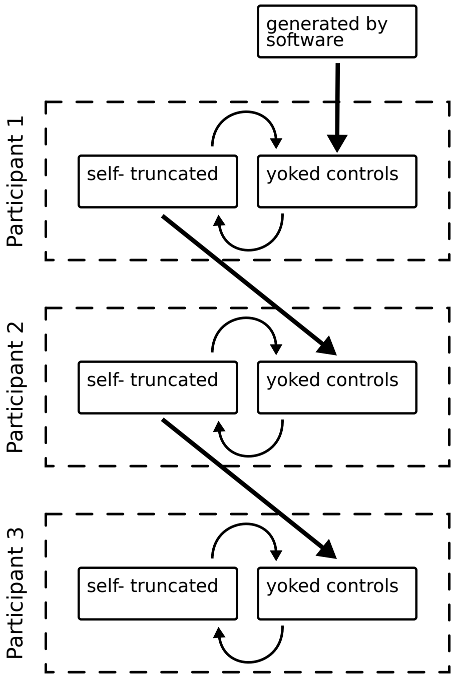


Figure 8. Serial yoked control design used for Experiment 2. Each participant received the self-truncated half of the samples from a preceding participant as his/her yoked control, while self-truncating the other half, which were in turn passed on to the next participant in the chain.

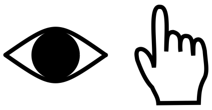


Figure 9. Icons announcing the type of trial to follow: observing an other-truncated sample (eye) or generating a self-truncated sample (hand).

trials, they were instructed that pressing the space bar would serve to sample more trait adjectives (one per keypress). Alternatively, they could truncate the sample by pressing "Enter". On yoked control trials, they observed the sample unfolding at a rate of 1 s per trait until it reached the size determined by the yoked partner's truncation decision. When sampling was complete, participants judged likeability and confidence in their judgment just as in Experiment 1. Participants were neither told how other-truncated samples had been generated nor that their own self-truncated samples were passed on to another participant, instructions merely discriminated between self-truncated sampling and observation of (yoked control) samples.

Analogous to Experiment 1, four partly overlapping population sets of 30 traits each were formed from the trait material. Again, two of these population sets were predominantly negative and two were positive and within both pairs of same-valence sets, one contained extreme and one contained moderate traits, overall. Each participant made 40 impression judgments, of which 20 were based on self-truncated samples (again with a maximal sample size of 16) and 20 predetermined samples. Within each subset of 20, samples were equally often drawn from all four population sets (i.e. five samples per set per sampling mode).

Participants. One hundred and fourteen participants (88 female) were recruited at

Table 10. Regression of Likeability Judgments on Population Set and Sampling Error

Predictor	M β (SD)	Consensus	t	df	p-value
Self-truncated trials					
Population average	.72 (.15)	100%	48.27	104	$< .001$
Sampling error	.34 (.16)	98%	22.24	104	$< .001$
Yoked controls trials					
Population average	.69 (.16)	99%	44.10	98	$< .001$
Sampling error	.31 (.16)	99%	19.66	98	$< .001$

Heidelberg University. Ninety-seven participants were students; their age ranged from 18 to 59 years ($M = 24.30$). Eight participants were excluded for always choosing the same sample size, and one for exhibiting strongly extended median response time for judgments (individual median response time more than four standard deviations above the mean of individual medians). From the remaining $N = 105$ participants, it was possible to form 99 yoked control pairs. For every computer workplace in the lab, the first participant could not have a preceding yoked partner, and was thus provided with computer-truncated samples (see Figure 8). Thus, these samples and the yoked controls' trials based on them (i.e. yoked controls' trials by the first participants in each workplace) had to be removed from the data set. Finally, 91 yoked pairs containing actual truncation decisions remained for the analysis. Additionally, we made sure that no participant had already participated in Experiment 1.

Results

Systematic Nature of Impression Judgments and Sensitivity to Sampling Input. We repeated the checks on the systematic nature of judgments similar to the first experiment. Impression judgments were again firmly predictable from population set averages and sampling error. Table 10 summarizes these results separately for self-truncated and yoked controls trials. Judgments were again highly sensitive to systematic (i.e. population-determined) and to random variation (sampling error) in the trait input. The mean individual coefficients of determination were $R^2 = .71$ (self-truncated trials) and $R^2 = .65$ (yoked controls trials).

Relation of Sample Size and Judgment Strength. Correlations $r_{ln(n),J}$ between likeability judgment strength J (in the correct direction) and the natural logarithm of sample size $ln(n)$ were clearly negative (Figure 10), both for self-truncated samples (mean $r_{ln(n),J} = -.30$, $SD = .28$, $t(104) = -11.05$, $p < .001$, 84% consensus) and for yoked controls (mean $r_{ln(n),J} = -.11$, $SD = .24$, $t(98) = -4.42$, $p < .001$, 67% consensus).

Since data for self-truncated and yoked controls' trials were ideally matched in participant pairs, we were able to analyze the degree of regressive shrinkage between yoked partners' judgments pair-wise. The mean correlation between $ln(n)$ and J decreased by mean $\Delta r_{ln(n),J} = .19$ ($SD = .27$, $t(90) = 6.75$, $p < .001$, 73% consensus). Forming difference scores between individual yoked judgments, we again observed the same regressive pattern: judgment strength J decreased for small samples and tended to increase for large samples when comparing individual judgments based on self-truncated versus yoked control trials (see Figure 11). The difference score ΔJ thus correlated positively with the natural logarithm of sample size (mean $r_{ln(n),\Delta J} = .14$, $SD = .22$, $t(90) = 6.04$, $p < .001$, 69% consensus), reflecting regressive shrinkage of the distinct

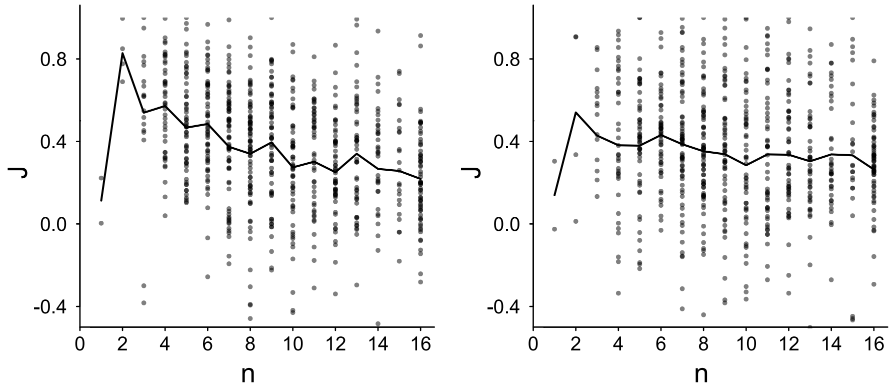


Figure 10. Correlations $r_{ln(n),J}$ between sample size n and judgment strength J for self-truncated samples (left hand side) and yoked control samples (right hand side). Dots represent individual mean values per condition and sample size, solid lines the averaged individual means per condition and sample size.

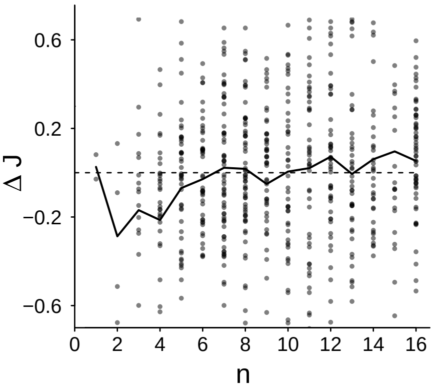


Figure 11. Changes in judgment strength ΔJ between self-truncated and yoked control samples between pairs of yoked partners plotted as a function of sample size n.

negative correlations $r_{ln(n),J}$ between sample size n and judgment strength J.

Diagnosticity and Expectedness. As in Experiment 1, we conducted a hierarchical multiple regression analysis of sample size n and judgment strength J, including as predictors the same indicators of diagnosticity: population set valence, extremity, and expectedness indicator sample word frequency.

The analysis of n as criterion revealed that sampling was truncated earlier, at lower n, when traits were drawn from a negative than when they were drawn from a positive set, and also when samples were drawn from extreme compared to moderate population sets (Table 11 and Figure 12). Additionally, samples of infrequent trait words were truncated earlier than samples of frequent trait words. However, as in Experiment 1, the β weight of the latter predictor (as distinguished from its substantial r) was negligible. The contribution of word frequency was absorbed by the other two correlated predictors (Table 12).

For both self-truncated and yoked controls, judgment strength J was higher (in the correct direction) when population sets were extreme and negative rather than moderate

Table 11. Regression of Self-Truncated $ln(n)$ on Indicators of Diagnosticity: Population Set Valence, Extremity, and Expectedness Indicator Sample Mean Word Frequency

Predictor	M r (SD)	Consensus	M β (SD)	Consensus	t	df	p-value
Population set extremity	-.12 (.25)	67%	-.12 (.27)	61%	-4.66	104	$< .001$
Population set valence	.14 (.27)	70%	.15 (.34)	64%	4.36	104	$< .001$
Sample mean word frequency	.07 (.24)	67%	0 (.33)	49%	-.07	104	.943

Note. Only data from self-truncated trials are included.

Table 12. Predictor Intercorrelations

Predictor	M r (SD)	Consensus
Population set extremity and valence	0 (0)	-
Population set extremity and sample mean word frequency	.15 (.18)	82%
Population set valence and sample mean word frequency	.56 (.18)	99%

Note. Only data from self-truncated trials are included.

and positive. Trait word frequency did not account for additional variance of judgment strength (see Table 13).

In a corresponding regression analysis of the correlation $r_{ln(n),J}$ between the natural logarithm of sample size n and judgment strength J as criterion, no noticeable influence of diagnosticity could be detected (Table 14).

Judgment Confidence. Subjective confidence c was again higher for smaller than for larger samples (Table 15), corroborating the enhanced strength of judgments for small samples. A comparison of $r_{ln(n),J}$ correlations for self-truncated and yoked controls' samples showed marked regressive shrinkage by mean $\Delta r_{ln(n),c} = .21$ ($SD = .36$, $t(90) = 5.50$, $p < .001$, 68% consensus).

Regarding diagnosticity, the confidence results did not completely mirror the findings for judgment strength J. Extreme population sets were associated with higher confidence and with stronger negative correlations between sample size and confidence in the judgment. However, population set valence and sample mean word frequency did not contribute to predicting the relation $r_{ln(n),c}$ between sample size and confidence.

Discussion

For Experiment 2 the entire stimulus materials were replaced and the design changed from a between-participants designs with two consecutive blocks to a repeated-measures design with self-truncated and other truncated trait samples alternating randomly

Table 13. Regression of Judgment Strength J on Indicators of Diagnosticity: Population Set Valence, Extremity, and Expectedness Indicator Sample Mean Word Frequency, Separately for Self-Truncated and Yoked Control Trials

Predictor	M r (SD)	Consensus	M β (SD)	Consensus	t	df	p-value
Self-truncated trials							
Population set extremity	.24 (.24)	88%	.24 (.25)	87%	9.93	104	$< .001$
Population set valence	-.14 (.37)	64%	-.14 (.42)	66%	-3.44	104	.001
Sample mean word frequency	-.02 (.29)	56%	.01 (.28)	56%	.34	104	.734
Yoked controls trials							
Population set extremity	.23 (.20)	86%	.21 (.22)	84%	9.30	98	$< .001$
Population set valence	-.08 (.36)	62%	-.10 (.41)	60%	-2.41	98	.018
Sample mean word frequency	.01 (.26)	51%	.03 (.28)	58%	1.18	98	.241

Table 14. Regression of the Correlation $r_{ln(n),J}$ Between Judgment Strength and Sample Size on Indicators of Diagnosticity: Population Set Valence, Extremity, and Expectedness Indicator Sample Mean Word Frequency, Separately for Self-Truncated and Yoked Control Trials

Predictor	β	Consensus	t	*df*	p-value
Self-truncated trials					
Population set extremity	.03	46%	.66	393	.511
Population set valence	.10	57%	1.96	393	.051
Yoked controls trials					
Population set extremity	.06	50%	1.12	363	.265
Population set valence	0	51%	.08	363	.934

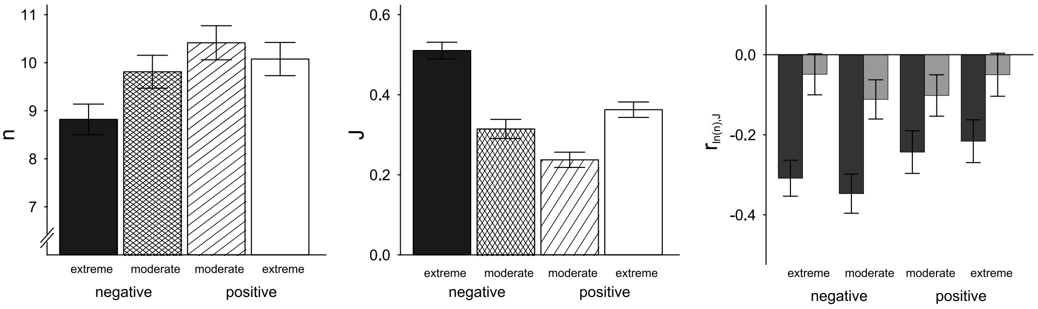


Figure 12. Sample size n, judgment strength J and the correlation $r_{ln(n),J}$ between the two variables as a function of the valence and extremity of population sets. Error bars indicate standard errors of the mean. The plot for $r_{ln(n),J}$ is broken down by sampling conditions; dark grey bars indicate self-truncated trials and light grey bars yoked controls' trials.

Table 15. Mean $r_{ln(n),J}$ Correlations Between the Natural Logarithm of Sample Size n and Self-Rated Confidence in Impression Judgments c

Sampling mode	M r (SD)	Consensus	t	*df*	p-value
Self-truncated	-.32 (.31)	82%	-10.82	104	$< .001$
Yoked controls	-.08 (.26)	63%	-3.02	98	.003

Table 16. Regression of Judgment Confidence c on Population Set Valence, Extremity, and Sample Mean Word Frequency

Predictor	M r (SD)	Consensus	M β (SD)	Consensus	t	*df*	p-value
Population set extremity	.17 (.24)	78%	.17 (.26)	76%	6.64	104	$< .001$
Population set valence	-.01 (.28)	53%	-.03 (.33)	56%	-1.02	104	.308
Sample mean word frequency	.03 (-23)	53%	.04 (.28)	55%	1.29	104	.201

Note. Only data from self-truncated trials are included.

Table 17. Regression of the Correlation $r_{ln(n),c}$ Between Confidence and Sample Size on Population Set Valence and Extremity

Predictor	β	Consensus	t	*df*	p-value
Population set extremity	.18	34%	3.56	393	$< .001$
Population set valence	-.03	40%	-.68	393	.496

Note. Only data from self-truncated trials are included. Results from yoked controls trials highly converge.

within participants. Despite these changes, an equivalent pattern of robust findings was obtained, testifying to the regularity and robustness of the interplay of Brunswikian and Thurstonian sampling.

Judgments remained highly sensitive to systematic (population parameters) and unsystematic (sampling error) sources of sample variation. Both sample size n and judgment strength J also reflected the diagnostic value of sampling input. Population set valence and extremity of the sampled input regularly determined participants' decisions to truncate the sample and the strength of their judgments. However, the expectedness parameter of word frequency within the samples no longer contributed to predicting participant behavior.

Participants' ratings strongly confirmed regressive shrinkage in judgments on perfectly identical samples of traits, which only differed in how the sample had been truncated: The relation between sample size and judgment strength declined considerably when passing the formerly self-truncated sample on to the subsequent yoked control. The confirmation of this regression phenomenon in the repeated-measures design of Experiment 2 provides cogent evidence for Thurstonian sampling effects, making alternative accounts in terms of differential response-scale usage between self-truncated and yoked controls trials highly improbable.

General Discussion

The two experiments and the simulation study reported in the present article provide strong convergent evidence about the highly regular nature of trait integration in person impression formation. In the context of a sample-based impression judgment paradigm, in which unknown target persons were solely described by samples of stimulus traits drawn at random from an experimentally controlled population set, we were able to rule out the influence of prior knowledge and memory-based target impressions (Hastie & Park, 1986). The resulting likeability judgments were highly predictable and strongly influenced by the judges' freedom to truncate the trait sampling process. Let us first summarize the evidence and the theoretical insights gained from the reported findings, before we discuss their implications in the context of the extant literature and related issues in current judgment and decision making research.

The results of our experiments highlight the extremely regular nature of sample-based impression judgments. Our mixed design with many judgment targets nested within individual participants revealed that judgments were highly sensitive to the average valence of the target's trait population set, to trait sampling error, diagnosticity, and the number of sampled traits. Predictions of impression ratings from the mere sampling input (i.e. the actuarial judgments; Dawes et al., 1989) resulted in determination coefficients ranging between $R^2 = .52$ and $R^2 = .71$. In other words, judgments were highly sensitive to Brunswikian sampling variance, that is, to the environmental input of trait stimuli provided on each judgment trial.

However, secondly, the manner in which the resulting impressions depend on the stimulus input diverged markedly from a naïve application of statistical sampling theory, which would suggest that judgment strength increases with the increasing size of a sample drawn from the same population set. Both prior empirical evidence (Fiedler & Kareev, 2006; Fiedler et al., 2010) and normative principles such as Bayesian updating imply that the same observed tendency of the sample toward mainly positive or mainly negative traits provides stronger evidence in larger than in smaller samples. And, indeed, prior research in the present sample-based judgment paradigm confirmed the validity of this implication when sample size was implemented as an independent variable, that is, when n was pre-determined by design, so that sample truncation was independent of the strength of evidence sampled so far. In contrast, when the truncation decision was drawn by the judges themselves, making n strongly dependent on the input and on primacy effects, a strong and consistent reversal was obtained: Smaller (self-truncated) samples led to stronger judgments than larger samples drawn from the same population. Simulation results corroborated these remarkable results and clarified that the negative correlation between judgment strength and samples size can be emulated by an objectively defined algorithm.

This strong and robust reversal, which can be predicted on normative grounds, has important theoretical and practical implications. Understanding the stronger evidence conveyed in small rather than large samples at a theoretical level requires going beyond the basic assumptions of a simple averaging model (Anderson, 1965; Ullrich et al., 2013), which presumes that likeability judgments reflect the average valence scale values of all observed stimulus traits. Despite its simplicity and usefulness as a comparison standard, the averaging rule does not account for truncation effects and, more importantly, it does not acknowledge the traits' informational value (i.e. diagnosticity) beyond mere valence in a given impression formation context.

The diagnosticity of sampled traits not only received a strong weight in impression judgments, but also facilitated early truncation of samples, leading to strong negative correlations between sample size and judgment strength. Negative traits were clearly more diagnostic than positive traits, although their absolute valence scale values were matched. Thus, diagnosticity is not an isolated feature of individual trait stimuli but a measure of a trait's impact on an integrative judgment exceeding the item-level valence.

Within the broader context of recent theorizing in judgment and decision making research, the present approach also emphasizes the need to go beyond the so-called "description-experience gap" (Hertwig et al., 2004; Hertwig & Pleskac, 2010). This major topic of recent research on sample-based decision making (Wulff et al., 2018) focuses on the contrast between decision options described (second-hand) in terms of numerically specified expectancies and probabilities and the (first-hand) experience with a sample of observations under uncertainty, from which expected values and probabilities must be inferred in an inductive-statistical process. The present findings demonstrate, however, that it is essential to further distinguish between different types of decisions by experience, depending on whether samples are self-truncated or predetermined externally.

The negative correlation of sample size and impression judgment strength for self-truncated samples has important practical implications as well. For instance, Wilson and Schooler's (1991) finding that consumer choices are met with higher satisfaction when reasoning about the choice was brief rather than extended may be understood as a special case of self-truncated choice. This finding is not at all in conflict with other evidence for higher decision quality with increasing amount of information, provided the size of information samples is determined externally.

Practical consequences and misunderstandings of the self-truncation effect can be expected because the reversal correlations between judgment strength and the size of self-truncated versus externally fixed samples is highly counterintuitive. In many applied domains, decision makers may hardly ever realize and memorize the corresponding truncation rule. Consumers barely notice whether they stopped sampling because they had gathered enough information or because of external constraints. Teachers hardly know how their samples of different students' performance were truncated, just like members of democratic decision groups do not typically include a note in the protocol on what determined the end of a deliberation process. Consequently, it is hard to distinguish whether distinct and clear-cut information is either a symptom of diligence and accuracy (if n is fixed) or a sign of cogent initial evidence (if sampling is self-truncated).

When impressions are made from conditionally truncated samples, on a trial basis (e.g. grading one particular student) we have a hard time distinguishing between the influence of chance (i.e. "exploitation of good luck"; Edwards, 1965) and the influence of an underlying trend. Thus, the negative relation between self-truncated sample size and impression strength is not a mere cognitive bias but is reflective of individuals' reactions to an insolvable statistical dilemma. Ignoring opportunities of "good luck" (i.e. small samples drawing an extremely clear picture) can be very costly, as we need to waste time and resources after we have already seen a clear hint on how to form the impression. As our analysis of impression data shows, exploiting such primacy opportunities does not prevent judges from making highly accurate judgments, coming close to the averaging benchmark (Anderson, 1965). In any context where samples cannot be extended infinitely, because judgments need to be completed in time, dynamic self-truncated sampling is a highly efficient and adaptive tool, as it exploits instances of initially clear-cut information.

Whether we consider the often exaggerated views provided by very small samples to be excessive or advantageous critically depends on the context. Small sample amplification of impressions is advantageous in contexts where the benefits of efficient trend-detection are high. In such contexts, yoked controls' judgments might be interpreted as too cautious or even obscuring the correct trends. In contrast, in other contexts that render the costs of exaggerating impressions and erroneous inferences very expensive, more cautious judgment strategies based on generally stronger evidence (thus, resulting in larger samples) are called for.

Last but not least, our sample-based judgment paradigm helps to illuminate the joint impact of two intertwined sources of information sampling. Person impressions were shown to be contingent not only on Brunswikian sampling of traits in the stimulus environments but also on Thurstonian sampling processes taking place within the judges' mind. In other words, it was shown that impressions are not exclusively determined by the trait input; they also reflect the oscillations that take place in the mind of the beholder.

Although impression judgments were highly predictable from, and greatly resembled an actuarial assessment of the Brunswikian sample input, they were also contingent on Thurstonian sampling of mental states, that is, on fluctuations in judges' mental preparedness to make a judgment. Thus, judges stopped sampling and finally made a judgment not only when the sample of traits was informative and diagnostic but also when they happened to be in a state of mind that enabled them to recognize and perhaps even overestimate the evidence in the sample.

To render this second sort of Thurstonian sample visible, we developed a yoked control design that allowed us to compare the original self-truncating judges' impressions with the impressions of yoked controls, who were provided with exactly the same trait samples, presented in the same format, and trait order, and who could therefore be expected to differ only in terms of Thurstonian oscillations of states of mind. Unlike self-truncating judges, yoked controls cannot be ideally prepared for a judgment at the very moment when the trait sample was complete. As a consequence of this misfit – the only distinctive feature of self-truncating judges and their yoked controls – the yoked participants' impression judgments were clearly regressive. Positive impressions were less positive and negative impressions were less negative than in the self-truncating conditions.

We elaborated on this methodology, showing that yoked control judges provided with their own pre-determined samples showed less regression than yoked controls provided with other judges' previously determined samples. This refined method allowed us to distinguish between temporal and interpersonal sources of Thurstonian sampling effects. Thurstonian misalignment of yoked partners most probably presents one source of disagreement between individuals' impressions when one self-truncated sample was passed on to another person, or even when the same individual re-evaluated their own impressions. Overall, yoked partners' regressive judgments were more moderate. In our experimental setting, yoked partners most probably disagreed when the self-truncated part of the pair came to a judgment of strong disliking after only a few pieces of information. However, yoked controls introduced their own individual deviations from the mere Brunswikian sampling input. As determination coefficients in explaining impression variance by actuarial (averaging) impressions showed, this did not serve to improve precision of judgments; yoked controls rather deviated further from the actuarial estimation.

Though the present article takes a strong theoretical focus, yoked control settings are also at the heart of many everyday social phenomena. Whenever someone tries to pass on information or an opinion from self-truncated information search, it is the opinions on topics supported by small information samples that have the largest potential for disagreement. In turn, strong impressions from self-truncated small samples are likely to drive the message passed on in a serial-reproduction process (Kashima, 2000). We also showed theoretically that the re-evaluation of one's own impression, judgment or opinion after examining the original sources underlying the judgment again might cause a tendency to be dissatisfied and to feel conflict with, or simply to revise one's very own initial impression, due to regressive shrinkage.

We believe that the presented findings and implications can lead to an improved understanding of person impressions. With respect to the ongoing debate on the restricted replicability and usability of psychological findings, we would like to emphasize once more the regularity and robustness of person impressions and the highly systematic pattern of a negative relation between sample size and the strength of judgments on self-truncated samples. Just as the evidence inherent in a pre-determined sample increases with sample size, the higher deviation of smaller samples creates the potential for the counterintuitive finding that the evidence inherent in a self-truncated sample is stronger when samples size is low rather than high.

References

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, *70*(4), 394–400.

Bernoulli, J. (1713). *Ars conjectandi: Opus posthumum.* Thurnisii.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674.

De Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, *7*(1), 1–68.

Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology*, *2*(2), 312–329.

Fiedler, K., & Kareev, Y. (2006). Does decision quality (always) increase with the size of information samples? some vicissitudes in applying the law of large numbers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 883–903.

Fiedler, K., Renn, S.-Y., & Kareev, Y. (2010). Mood and judgments based on sequential sampling. *Journal of Behavioral Decision Making*, *23*(5), 483–495.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889–906.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

Galton, F. (1907). Vox populi. *Nature*, *75*(1949), 450–451.

Gidron, D., Koehler, D. J., & Tversky, A. (1993). Implicit quantification of personality traits. *Personality and Social Psychology Bulletin*, *19*(5), 594–604.

Hadar, L., & Fox, C. R. (2009). Information asymmetry in decision from description versus decision from experience. *Judgment and Decision Making*, *4*(4), 317–325.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review*, *93*(3), 258–268.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539.

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, *115*(2), 225–237.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, *104*(2), 344–366.

Kahan, J. P., Rapoport, A., & Jones, L. V. (1967). Decision making in a sequential search task. *Perception Psychophysics*, *2*(8), 374–376.

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review*, *107*(2), 397–402.

Kashima, Y. (2000). Maintaining cultural stereotypes in the serial reproduction of narratives. *Personality and Social Psychology Bulletin*, *26*(5), , 594–604.

Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation*, *15*, 192–238.

Koch, A., Alves, H., Krüger, T., & Unkelbach, C. (2016). A general valence asymmetry in similarity: Good is more alike than bad. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *42*(8), 1171–1192.

Leibniz Institut Für Deutsche Sprache. (2014). Datenbank für gesprochenes Deutsch [spoken german database].

Leibniz Institut Für Deutsche Sprache. (2017). Deutsches Referenzkorpus [german reference corpus].

Parducci, A., & Perrett, L. F. (1971). Category rating scales: Effects of relative spacing and frequency of stimulus values. *Journal of Experimental Psychology*, *89*(2), 427–452.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning*, *Memory*(3), 534–552.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*(1), 29–46.

Prager, J., Krueger, J. I., & Fiedler, K. (2018). Towards a deeper understanding of impression formation-new insights gained from a cognitive-ecological perspective. *Journal of Personality and Social Psychology*, *115*(3), 379–397.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, *86*(1), 61–79.

Savage, L. J. (1954). *The foundations of statistics.* Wiley.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology*, *52*(4), 689–699.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271–295.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286.

Ullrich, J., Krueger, J. I., Brod, A., & Groschupf, F. (2013). More is not less: Greater information quantity does not diminish liking. *Journal of Personality and Social Psychology*, *105*(6), 909–920.

Unkelbach, C., Fiedler, K., Bayer, M., Stegmüller, M., & Danner, D. (2008). Why positive information is processed faster: The density hypothesis. *Journal of Personality and Social Psychology*, *95*(1), 36–49.

Võ, M. L. H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin affective word list reloaded (BAWL-R). *Behavior Research Methods*, *41*(2), 534–538.

Wald, A. (1947). *Sequential analysis.* John Wiley.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, *60*(2), 181–192.

Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A metaanalytic review of two modes of learning and the description-experience gap. *Psychological Bulletin*, *144*(2), 140–176.

Appendix A. Experiment 1: Construction of the Four Trait Population Sets

Fifty-seven trait adjectives of the Berlin Affective Word List – Reloaded (BAWL-R) by Võ et al. (2009) made up the pool of potential stimuli. Four potentially overlapping population sets were selected from the overall stimulus pool. The aim was to generate almost symmetrical population sets. Therefore, the BAWL-R scale was split into six equally wide intervals. As can be seen from Table A1, certain frequencies in these valence intervals were considered.

Table A1. Fixed Frequencies for Generating the Population Sets. Adjectives of Each Population Set Were Drawn at Random, However The Proportions of Valence Intervals Were Enforced.

	BAWL-R valence interval					
Population set	-3 : -2	-2 : -1	-1 : 0	0 : 1	1 : 2	2 : 3
Extremely negative	4	12	8	4	1	1
Moderately negative	2	10	8	7	2	1
Moderately positive	1	2	7	11	7	2
Extremely positive	1	1	4	11	11	2

This selection lead to almost symmetrical properties of the two extreme and the two moderate sets (cf. Table A2).

Table A2. Proportion of Positive Traits p_+ and Statistics for BAWL-R Valence Norms (Scaled to a Range of -1 : 1) for Each of the Four Population Sets.

Population set	p_+	*M*	*SD*	Skewness
Extremely negative	.20	-.26	.38	.88
Moderately negative	.33	-.15	.41	.39
Moderately positive	.67	.16	.38	-.56
Extremely positive	.80	.26	.37	-.97

Appendix B. Experiment 2: Consruction of the Four Trait Population Sets

Seventy-one trait adjectives scaled in the pretest formed the original pool of potential stimuli. The formation of the four population sets followed similar principles as in Experiment 1. The frequency table for the population sets is shown in Table B1 and the properties of the resulting set in Table B2.

Table A3. Trait Adjectives Used in Experiment 1 With their English Translation and Their BAWL-R Mean Valence (Scaled from Very Negative -3 to Very Positive 3; Võ et al., 2009). The Columns on the Right Hand Side Indicate the Usage (1) or Non-Usage (0) of the Trait for Each of the Four Population Sets Indicated by Their Proportion of Positive Valent Traits.

Original German	English translation	BAWL-R valence	$p_+ = .20$	$p_+ = .33$	$p_+ = .67$	$p_+ = .80$
herzlos	heartless	-2.5	0	1	0	0
verlogen	mendacious	-2.3	1	0	1	0
humorlos	humorless	-2.1	1	0	0	1
unfair	unfair	-2	1	0	0	0
dekadent	decadent	-2	1	0	0	0
brutal	harsh	-2	0	1	0	0
launisch	moody	-1.9	1	1	0	0
gemein	mean	-1.9	1	1	0	0
boshaft	mischieveous	-1.9	1	1	0	0
gierig	greedy	-1.5	1	1	1	1
stur	stubborn	-1.5	1	1	0	0
labil	labile	-1.5	1	1	0	0
mutlos	discouraged	-1.4	1	1	1	0
primitiv	primitive	-1.4	1	0	0	0
altklug	precocious	-1.1	1	1	0	0
eitel	vein	-1.1	1	1	0	0
passiv	passive	-1.1	1	1	0	0
derb	coarse	-1.1	1	0	0	0
wortkarg	taciturn	-0.9	1	1	1	1
naiv	naive	-0.8	1	0	0	0
laut	noisy	-0.8	1	1	1	1
defensiv	defensive	-0.6	1	1	1	0
unnahbar	unapproachable	-0.6	0	1	0	0
listig	cunning	-0.4	1	0	1	1
albern	ridiculous	-0.2	1	1	1	1
forsch	outspoken	-0.2	1	1	1	0
redselig	talkative	-0.1	0	1	0	0
schlicht	plain	0	1	1	1	0
stoisch	stoical	0.1	0	1	1	1
sparsam	thrifty	0.2	0	1	1	1
still	silent	0.6	1	1	1	1
verwegen	audacious	0.6	1	0	1	1
ruhig	calm	0.6	0	1	1	1
sachlich	factual	0.6	0	0	1	1
eifrig	eager	0.8	1	1	1	1
sensibel	sensitive	0.8	1	1	1	1
vornehm	genteel	0.9	0	0	1	1
strebsam	ambitious	0.9	0	0	1	1
agil	agile	1	0	1	1	1
sanft	gently	1.3	1	0	0	1
liberal	liberal	1.3	0	0	0	1
spontan	spontaneous	1.4	0	0	1	1
schlau	clever	1.6	0	1	1	1
aktiv	active	1.6	0	0	1	1
pfiffig	smart	1.6	0	0	1	0
flexibel	flexible	1.7	0	1	1	1
tapfer	brave	1.7	0	0	1	0
munter	cheerful	1.7	0	0	0	1
heiter	humorous	1.8	0	0	0	1
nett	kind	1.8	0	0	0	1
taktvoll	tactful	1.9	0	0	1	1

Table B1. Fixed frequencies for Generating the Population Sets of the Second Experiment.

	Prestudy valence interval					
Population set	-1 : -.66	-.66 : -.33	-.33 : 0	0 : .33	.33 : .66	.66 : 1
extremely negative	5	10	9	4	1	1
moderately negative	2	10	8	7	2	1
moderately positive	1	2	7	8	10	2
extremely positive	1	1	4	9	10	5

Table B2. Proportion of Positive Traits p_+ and Statistics for BAWL-R Valence Norms (Scaled to a Range of -1 : 1) for Each of the Four Population Sets.

Population set	p_+	M	SD	Skewness
extremely negative	.20	-.28	.38	.80
moderately negative	.33	-.18	.41	.58
moderately positive	.67	.19	.42	-.45
extremely positive	.80	.29	.42	-.88

Table B3. Trait Adjectives Used in Experiment 2 With their English Translation and Their Mean Valence (Scaled from Very Negative -1 to Very Positive +1) Assessed in a Pretest to Experiment 2. The Columns on the Right Hand Side Indicate the Usage (1) or Non-Usage (0) of the Trait for Each of the Four Population Sets Indicated by Their Proportion of Positive Valent Traits.

Original German	English translation	Valence	$p_+ = .20$	$p_+ = .33$	$p_+ = .67$	$p_+ = .80$
herzlos	heartless	-0.85	1	0	0	0
hochnäsig	arrogant	-0.80	1	0	0	1
ignorant	ignorant	-0.77	0	1	0	0
aggressiv	aggressive	-0.74	0	0	1	0
arrogant	conceited	-0.72	1	1	0	0
feindselig	hostile	-0.71	1	0	0	0
manipulativ	manipulative	-0.68	1	0	0	0
nervig	annoying	-0.65	0	1	0	0
oberflächlich	superficial	-0.65	0	1	0	1
aufgesetzt	hypocritical	-0.62	0	1	0	0
ichbezogen	egocentric	-0.60	0	0	1	0
dreist	audacious	-0.60	1	0	0	0
pessimistisch	pessimistic	-0.58	1	0	0	0
vulgär	coarse	-0.58	1	1	0	0
künstlich	artificial	-0.57	1	1	0	0
aufbrausend	irascible	-0.53	0	1	0	0
faul	lazy	-0.53	1	1	0	0
stur	stubborn	-0.51	1	1	0	0
anstrengend	exhausting	-0.46	1	0	1	0
dominant	dominant	-0.41	1	0	0	0
übertrieben	exaggerated	-0.38	0	1	0	0
verschlossen	withdrawn	-0.35	1	0	0	0
hässlich	ugly	-0.35	1	0	0	0
schwatzhaft	talkative	-0.34	0	1	0	0
bestimmend	determining	-0.28	1	1	1	0
überdreht	overwrought	-0.26	1	1	1	0
zerstreut	absentminded	-0.24	1	1	1	1
langsam	slow	-0.23	1	1	1	1
kränkbar	easily offended	-0.21	1	1	0	1
chaotisch	messy	-0.18	1	1	0	0
introvertiert	introverted	-0.13	1	1	1	0
impulsiv	impulsive	-0.13	1	0	1	1
schüchtern	shy	-0.01	1	1	1	0
still	silent	0.04	1	0	0	1
verschwiegen	discreet	0.05	0	1	0	1
alternativ	alternative	0.06	0	0	1	1
detailverliebt	attentive to detail	0.06	0	1	0	0
perfektionistisch	perfectionist	0.08	0	0	1	1
verrückt	crazy	0.10	1	1	1	1
zurückhaltend	reserved	0.12	0	1	1	1
vorsichtig	cautious	0.20	0	1	0	1
rational	rational	0.22	1	1	1	0
selbstironisch	self-ironic	0.31	0	0	1	0
kumpelhaft	back-slapping	0.32	1	1	1	1
ruhig	calm	0.33	0	0	1	1
bedacht	considered	0.40	0	0	1	0
spontan	spontaneous	0.47	1	0	1	0
stabil	stable	0.50	0	0	1	1
neugierig	curious	0.50	0	0	0	1
empathisch	empathetic	0.53	0	0	1	0
gründlich	thorough	0.55	0	0	1	1
gesellig	sociable	0.55	0	1	0	1
inspirierend	inspiring	0.58	0	1	0	1
aktiv	active	0.59	0	0	1	0
gelassen	serene	0.60	0	0	0	1
warm	warm	0.60	0	0	0	1
offen	amenable	0.61	0	0	0	1
intelligent	intelligent	0.62	0	0	1	0
zärtlich	tender	0.62	0	0	1	1
authentisch	genuine	0.62	0	0	0	1
natürlich	natural	0.65	0	0	1	0
entspannt	relaxed	0.65	0	0	1	0
optimistisch	optimistic	0.71	0	0	0	1
kooperativ	cooperative	0.72	1	0	0	0
herzlich	sincere	0.75	0	0	0	1
mitfühlend	compassionate	0.76	0	1	0	0
treu	loyal	0.76	0	0	0	1

Appendix C

Small-Group Homogeneity: A Crucial Ingredient to Inter-Group Sampling and Impression Formation.

Small-Group Homogeneity: A Crucial Ingredient to Inter-Group Sampling and Impression Formation

Johannes Prager and Klaus Fiedler

Heidelberg University

ABSTRACT

Applying a recently developed framework for the study of sample-based individual person impressions to the level of group impressions resulted in convergent evidence for a refined but highly robust judgment process. Group impressions were sensitive both to systematic variance between distinct population sets from which trait samples were drawn and to the specific sampled traits. However, impressions did not merely follow a simple averaging rule applied to the likeability scale values of the sampled trait stimuli. Rather, the function relating group impressions to stimulus traits was subject to two distinct moderating influences, the diagnosticity of traits and the amplifying impact of early self-truncation. Three indices of trait diagnosticity – negative valence, extremity, and distance to other traits in a density framework – jointly determined not only the final impressions but also the decision to truncate the sampling process. When trait samples carried negative and extreme information and when traits within the sample were dense (i.e. the distance between traits was low), they triggered polarized impression judgments, high perceived within-sample homogeneity and early truncation decisions. Granting that out-group judgments typically rely on smaller samples than in-group judgments, our sampling approach can account for essential biases of inter-group judgment: out-group homogeneity, out-group polarization and (because negative traits are more diagnostic) out-group derogation.

KEYWORDS

out-group homogeneity, out-group polarization, self-truncated sampling

Out-group homogeneity (or relative in-group heterogeneity) is a classical and intensely discussed finding in social cognition and inter-group research. It refers to the unequal mental representation of groups we are part of (in-groups) and of groups we do not belong to (out-group). Specifically, the variability between individuals and target behaviors is perceived to be lower for out-groups such as a rival university, another age group or a foreign nation than for in-groups such as one's own university, age group, or nationality (Linville et al., 1989; Quattrone & Jones, 1980).

Theoretical explanations of this asymmetry vary on a continuum, one pole of which emphasizes structural causes in the environment whereas the other pole emphasizes motives and conflicts within the individual. Typical structural causes include larger experience samples (Kareev et al., 2002; Konovalova & Le Mens, 2020; Linville & Fischer,

Johannes Prager, Department of Psychology, Heidelberg University, johannes.prager@psychologie.uni-heidelberg.de

Klaus Fiedler, Department of Psychology, Heidelberg University

This manuscript is currently unpublished. It was submitted for publication November 29, 2021.

The work underlying the present article was supported by a Grant provided by the Deutsche Forschungsgemeinschaft (Fi 294/29-1) to Klaus Fiedler. Helpful comments by Gaël Le Mens and Linda McCaughey on a draft of this article are gratefully acknowledged.

1993; Linville et al., 1989) and more detailed knowledge (Park & Rothbart, 1982) of in-groups compared to out-groups. In contrast, motivational accounts stress the desire to develop a positive in-group identity (Simon & Brown, 1987) along with optimal distinctiveness (Brewer, 1993) and the familiarity advantage of closer individuals. While there was never any doubt that motivational biases nourished by real conflicts, group-related emotions or xenophobia are sufficient to trigger inter-group biases, a controversial question is whether they represent necessary conditions. Proponents of cognitive ecological theory approaches have recently pointed out that biased judgments and decisions can originate in completely unbiased intrapsychic processes (Denrell & Le Mens, 2007; Fazio et al., 2004; Fiedler, 2000; Fiedler & Wänke, 2009) embedded in perfectly adaptive behavior. Even when all stimuli are processed the same way, whether they are positive or negative in valence and related to in-groups or out-groups, the resulting judgments or evaluations can exhibit a systematic bias, simply because the stimulus environment provides the same mechanism with unequal samples of in-group or out-group related information. One obvious source of inequality, which is the focus of the present research, is sample size. Because our own group membership creates more opportunities to observe behaviors of other in-group members at an enhanced rate and reduced distance, in comparison to out-group members' behaviors, it seems self-evident that we are exposed to larger samples of in-group than out-group information (Quattrone & Jones, 1980).

Sample Size and Inter-Group Relations

The aim of the present investigation is to demonstrate that this characteristic difference in sample size offers a comprehensive account of out-group homogeneity. Computer simulations and a series of two pilot studies and four main experiments not only provide convergent evidence for a strong and regular out-group homogeneity effect. They also demonstrate that the same mechanism that produces out-group homogeneity also produces the other major phenomenon of inter-group research, namely, out-group derogation. Thus, the mechanism propagated in the present research can explain why (out-)groups described by smaller samples not only appear more homogeneous but also more negative than (in-)groups described by larger samples. We hasten to add that our evidence specifies a sufficient condition for inter-group biases. It need not reflect a necessary condition as it cannot exclude that other causal mechanisms may also produce inter-group biases.

A review of previous research suggests two different explanations of why small sample size may be at the heart of (out-)group homogeneity. The first explanation can be derived from statistical sampling theory, as formalized in Linville et al.'s (1989) seminal work. The loss of one degree of freedom in the calculation of a sample variance implies a stronger variance reduction for small samples (when $n - 1$ is markedly lower than n) than for large samples (when $n - 1$ approximates n). The resulting decrease in actual variance with decreasing n shown by Kareev et al. (2002) to account for a sizeable part of judgment biases.

In contrast to this formal statistical proof, the second explanation attributes the reduced variance of smaller samples to a "hot-stove" effect (Denrell & March, 2001). Assuming that adaptive agents follow Thorndike's (1927) law of effect, repeating pleasant and stopping unpleasant behaviors, they will under specifiable conditions (Denrell, 2005; Denrell & Le Mens, 2007) truncate sampling from negative sources while continuing to sample from pleasant sources. Consequently, they forego to correct for initial negativity effects (i.e., persistent and markedly negative judgments of small-sample targets) and only have a chance to correct for initially positive impressions (leading to larger samples of moderately positive stimuli). Thus, unlike the statistical sampling rule, this explanation attributes the sample-size effect to the agent's hedonic information search, which is neither irrational nor driven by any a-priori bias against some stigmatized target (Brockbank et al., in press; Le Mens & Denrell, 2011).

Self-Truncation Effects

The mechanism proposed in the present research is categorically different from both of these earlier approaches. It can neither be reduced to the statistical $(n-1)/n$ correction that underlies Linville's work, nor does it reflect a hot-stove effect as proposed by Denrell and Le Mens or Fazio and colleagues. Rather, the present account is based on recent findings uncovered and analyzed in our own research on impressions of individual targets based on samples of traits (Prager, Krueger & Fiedler, 2018; Prager & Fiedler, in press). Across a series of experiments, we persistently found that the strength of impressions J increased with increasing samples of n traits, but only when sample size was determined experimentally as an independent variable. When however, participants in a self-truncated sampling condition could themselves determine sampling when they felt to be ready for an impression judgment, making n a dependent variable, judgment strength J bore a clearly negative correlation to sample size n.

Note that this clear-cut reversal from positive correlations (with experimenter-determined sampling) to negative correlations (with self-truncated sampling) is fully consistent with statistical sampling principles. Although sample statistics indeed approximate population parameters as n increases, so that large samples more likely reflect existing population trends than smaller samples (De Finetti, 1937), Bernoulli's (1713) law of large numbers leads to new implications when samples are self-truncated. Self-truncated sampling produces small samples when the first few items exhibit a strong, presumably stable, conflict-free and consistent trend, whereas they become large only when no such stochastic primacy effect allows for early truncation. As a sample's deviation from population parameters becomes more probable with decreasing sample size, stronger stochastic primacy effects can be exploited when samples are small rather than large.

Diagnosticity Amplifies Self-Truncation Effects.

However, the basic self-truncation effect – overlooked and counter-intuitive as it may appear – is but one part of the story underlying the mechanism we are propagating. Equally important as the negative correlation it implies between n (sample size) and J (judgment strength) is the moderating impact of diagnosticity. Not every primacy trend (i.e. evidence from the first few items) is equally likely to cause truncation. Rather, the likelihood to truncate a sample increases markedly with the diagnosticity of the stimuli sampled so far. Early sampling is clearly more likely when the first few items in a sample are high rather than low in diagnosticity. Traits are more diagnostic if they are extreme rather than moderate and if there are negative rather than positive (Prager et al., 2018; Rothbart & Park, 1986). Moreover, diagnosticity comes to interact with the "big two", such that traits referring to negative morality and positive ability are more diagnostic than traits referring to positive morality and negative ability (Fiske et al., 2007; Reeder & Brewer, 1979; Skowronski & Carlston, 1987). Thus, the systematic tendency of small trait samples to solicit strong judgments, negatively correlated with sample size, is an increasing function of all these diagnosticity functions.

Let us take this opportunity to be explicit about the definition and theoretical meaning of diagnosticity. Whereas positive versus negative valence, extremity, and reference to the big two are semantic features of individual traits, a trait's diagnosticity quantifies its impact on a sequentially updated impression. In Bayesian odds notation, the likelihood ratio represents diagnosticity, quantifying the degree to which an added trait renders one final (e.g., positive) impression more likely than its reverse (e.g., negative). To illustrate this point, Prager et al. (2018) found that negative traits exerted stronger impact on resulting impressions than positive trait words of the same absolute valence scale value.

Density Model

The density model (Unkelbach et al., 2008) offers a deeper understanding of the cognitive underpinnings of the diagnosticity concept and the way it moderates cognitive performance. Positive stimuli are closer to each other and more densely interconnected in associative memory than negative stimuli, which are more distinct in meaning and less overlapping. For example, if someone is polite, they are also very likely friendly and punctual and reliable and tactful. In contrast, if someone is dishonest, we can hardly infer that they are offensive, brutal, depressed, or resentful. Thus, whereas positive person attributes form interconnected clusters, producing integrative halo effects (Unkelbach et al., 2008), negative attributes denote more separable properties. As a consequence, positive person attributes are common and redundant and positive words are used more frequently, whereas negative attributes are distinct and negative words are less frequent. Because of their more specific, less overlapping meaning, negative words are more diverse. The lexicon contains more distinct negative than positive verbs and adjectives.

At the behavioral level, the positive priming effects are stronger and positive words can be recognized and positive person attributes verified faster than negative stimuli (Unkelbach et al., 2008), simply because high-density clusters of positive stimuli in associative memory allow for a good deal of parallel processing. For the same reason, though, singular positive words or person attributes add little to the semantic meaning of the other positive items in the cluster, and a higher rate of positive evidence is required to confirm a positive judgment (Gidron et al., 1993; Rothbart & Park, 1986). Recognizing and confirming negative words and attributes takes longer but exerts stronger impact on evaluative judgments. Negative valence is thus a major determinant of diagnosticity, and the density model offers a mental account (i.e., the unequal distance and overlap or positive and negative stimuli in memory) and a highly useful measure of diagnosticity (i.e., the average distance from the remaining stimuli in a set, measured through multidimensional scaling; see Koch et al. 2016).

Yet, negative valence is by no means the only determinant of diagnosticity. Pitting diagnosticity against valence, it has been shown that those exceptional negative stimuli that bear low distances to others behave like high-density (non-diagnostic) stimuli and exceptional positive stimuli with high distances to others behave like low-density (diagnostic) stimuli (Unkelbach et al., 2008). Because extreme stimuli are also more distant from other stimuli, they were found by Prager et al. (2018) to exert stronger influence on growing impressions (i.e., to be more diagnostic) than moderate stimuli.

Drawing on the density model as a conceptual framework and as a methodological tool, we therefore expect stronger self-truncation effects resulting in stronger impression judgments for negative than positive traits, and for extreme than moderate traits. We do not expect that self-truncation and group impressions will be moderated by those aspects of the density model that are unrelated to diagnosticity, such as the frequency of occurrence of positive and negative terms in a large language corpus.[1]

Preview and Predictions

To lay out the theoretical expectations and empirical hypotheses to be tested below, we expect the typical self-truncation effect (i.e., negative correlation between n and J) to carry over from individual to group impressions. This basic prediction is not at all trivial. Because every trait in the group impression formation task refers to a different individual, entitativity is higher for individuals than for groups as impression targets (Campbell, 1958; Yzerbyt et al., 2000). It is therefore possible that weaker entitativity and consistency constraints moderate the impact of sample size on group impression judgments and reduce the influence of prior knowledge about existing social groups. To support this premise, we manipulate familiar versus neutral group labels to rule out prior group knowledge as an inhibiting condition and to highlight the functional equivalence of unlabeled samples of n people and existing groups with a familiar name. Note that, if anything, reduced entitativity implies higher independence of traits in a sample, making groups particularly prone to truncation effects.

We demonstrate that impressions resulting from self-truncated trait sampling exhibit strong and persistent homogeneity effects such that small samples of low variance are truncated earlier, leading to more polarized judgments than large samples. We regularly observe negative correlations $r_{n,J}$ between sample size and judgment strength as well as negative correlations $r_{n,H}$ between sample size and (two measures of) homogeneity within most individual participants (computed across judgment trials). Both $r_{n,J}$ and $r_{n,H}$ become stronger with two measures of diagnosticity, that is, when traits are extreme rather than moderate and when traits are negative rather than positive. Moreover, because negative valence is a chief determinant of diagnosticity, homogeneity (and polarization) come along with devaluation. Impressions tend to be most homogeneous and polarized when negative valence renders traits most diagnostic.

Beware of the conditional direction of the mechanistic account we are proposing. We postulate self-truncation as a sufficient rather than as a necessary causal condition of OHE. Thus, our theoretical argument is not that self-truncation effects (i.e., negative $r_{n,J}$ and $r_{n,H}$ correlations) underlie all manifestations of OHE. Our argument is rather that small sample size alone is sufficient to produce homogeneity, polarization, and devaluation. To the extent that out-group samples are smaller than in-group-referent samples, our causal mechanism predicts self-truncation to produce the depicted inter-group biases. Note, however, that the same theoretical argument applies to small samples of all kinds, not just out-groups, and it does not apply to contexts in which large samples are available about prominent or very large out-groups (Simon & Brown, 1987). Logically, our argument does not preclude that other causal influences may obscure or overshadow the predicted self-truncation effect.

Finally, this outline should reveal what is theoretically novel and original about the present approach, which has not been anticipated in previous inter-group research. As mentioned at the outset, the depicted self-truncation mechanism can be neither reduced to the statistical $(n-1)/n$ argument that motivated Linville's (1989) approach, which cannot account for the persistent negative correlations $r_{n,J}$ and $r_{n,H}$. Nor can it be considered a special case of a hot-stove effect leading to a one-sided negativity bias,

[1] Whether a trait word appears once or ten times per million may also be considered a linguistic measure of diagnosticity. It is however hardly relevant for group impression updating. Thus, frequency of words in the lexicon is less relevant for impression judgments than the distinctness and distance of added traits. Later in the results section we shall further elaborate on this consideration.

because self-truncation effects are not restricted to negativity. We also specify conditions under which self-truncation causes positivity biases when diagnosticity goes beyond valence. Moreover, our sample-based judgment approach offers a sensible account for both prominent inter-group biases, out-group homogeneity and out-group derogation, and for the co-occurrence of both major biases. Thus, although self-truncation effects are not peculiar to groups but have been already demonstrated for judgments of individual targets, placing them in the context of inter-group theorizing yields a variety of novel insights, about inter-group biases and downstream consequences of self-truncation, and also a new theoretical focus on truncation triggered by homogeneity.

In the next section, we present formal algorithms and computer simulation approaches to underline the generality and the logical cogency of the self-truncation mechanism. Then, we report two pre-studies supposed to control for prior group knowledge, to highlight the viability of the assumption that samples of n items behave like groups of n members, and to establish the density model as a theoretical framework for understanding the notion of diagnosticity. The remainder of this article is then devoted to a series of four experiments, in which participants judge the homogeneity and likeability of groups described by a sample of traits. Independent variables include the diagnosticity (valence and extremity) of different sets of traits from which the stimulus traits are sampled, frequency of traits, average distance of traits in a set, and prior group knowledge. Sample size depends on participants' self-truncation decisions and constitutes a major mediating variable. Judgments of group likeability, two measures of homogeneity, and various derived measures, such as $r_{n,J}$ and $r_{n,H}$, function as dependent variables.

Convergent Simulation of Sample Truncation Algorithms

Although the decrease in homogeneity and impression strength through self-truncation is novel and counter-intuitive at first sight, it does not reflect a strange, far-fetched empirical phenomenon. It is rather derivable on theoretical grounds, reflecting a corollary of Bernoulli's (1713) law of large numbers. Because distributions of sample statistics – not only of means but also of samples' variance, skewness, and curtosis – are more dispersed for small than large samples, truncation can cause more inflation of a sampled trend when early truncation renders samples small. A glance at the statistical underpinnings reveals that the sign of $r_{n,J}$ (correlation between self-truncated sample size and impression strength) and $r_{n,H}$ (between sample size and perceived within-group homogeneity) is regularly negative under a wide range of conditions.

Indeed, different stopping rules converge in producing both out-group homogeneity (negative correlation $r_{n,H}$ of sample size and perceived sample homogeneity) and out-group polarization, (negative correlation $r_{n,J}$ of sample size and impression strength), which can be expected to be particularly strong for negative impressions, due to enhanced diagnosticity (producing out-group derogation). Thus, whether the truncation threshold decreases with sample size n or uses a fixed criterion independently of n, simulated self-truncation effects converge in producing out-group homogeneity and out-group polarization (i.e., negative $r_{n,H}$ and $r_{n,J}$). Later in this article, four group-impression experiments will support the simulation effects.

Readers who are not interested in formal notation may simply skip the next section. To illustrate the normative constraints imposed on impression formation from dichotomous samples, let us assume for simplicity that each trait in a sample characterizes a group member as either "likeable" or "unlikeable", so that group impression judgments amount to estimating the probability $p(\text{"}\textit{likeable}\text{"})$ of encountering likeable group mem-

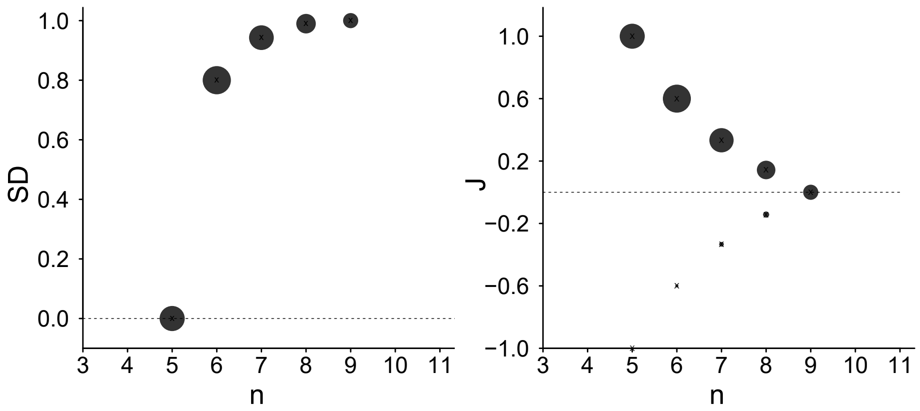


Figure 1. Simulated sample SD (inverse measure of homogeneity; left chart) and judgment strength J (right chart) as a function of n, the size of samples truncated according to Haldane's (1945) rule, assuming a true likeability probability of $p(\text{"}likeable\text{"}) = .75$ and a threshold of $t = 5$ (compromise between liberal and conservative strategy). Estimated impression strength is $J = 2 * \frac{k-1}{n-1} - 1$ when $k = t$ (or $J = 2 * \frac{n-k-1}{n-1} - 1$ when $n - k = t$).

bers. Given no prior knowledge of $p(\text{"}likeable\text{"})$, truncation decisions cannot rely on group's mean exceeding a threshold, but must rely on the second moment, namely on the variance of likeable versus unlikeable group members. Accordingly, a reasonable stopping rule is to cease sampling when group impression updates settle on a stable judgment, which hardly varies with additional traits/members.

Now, granting that low variability (homogeneity) triggers truncation, the question is whether the stopping criterion should decrease with increasing sample size n (as in statistical significance testing, when the standard error $se = SD/\sqrt{n}$ decreases with $\sqrt{n}$) or remains constant with increasing n (as in Bayesian updating). The first case of an n-dependent stopping rule was formalized by an algorithm suggested by Haldane (1945). The second case of an n-independent Bayesian stopping rule relies on updating a beta-distribution. In either case, the simulation results support stronger homogeneity and polarization for smaller than for larger samples.

Haldane's Labor-Saving Sampling Method

Assuming that judges, like statisticians, are sensitive to $\sqrt{n}$ Haldane (1945) proposed the standard error se as a variable threshold, calling for sample truncation when the number of "likeable" k traits (or the complementary number $n\breve{}k$ of "not likeable" traits) obtained in a sample of n exceeds an a-priori chosen threshold t. Figure 1 shows that this stopping rule implies that sample heterogeneity (conceived as the standard deviation $SD = se_t * \sqrt{n}$ increases (left chart) and impression strength J decreases with increasing sample size n (right chart). Note that a positive J reflects a deviation of judgments from .5 in the correct direction.

Belief-Updating Using the Beta-Distribution in Sample-Based Impressions

A Bayesian approach, in contrast, renders the stopping rule independent of $\sqrt{n}$.In Bayesian calculus, the posterior odds of, say, *p("likeable" | all sampled traits)/p("unlikeable" | all sampled traits)* equals the prior odds (set to 1/1 for uni-

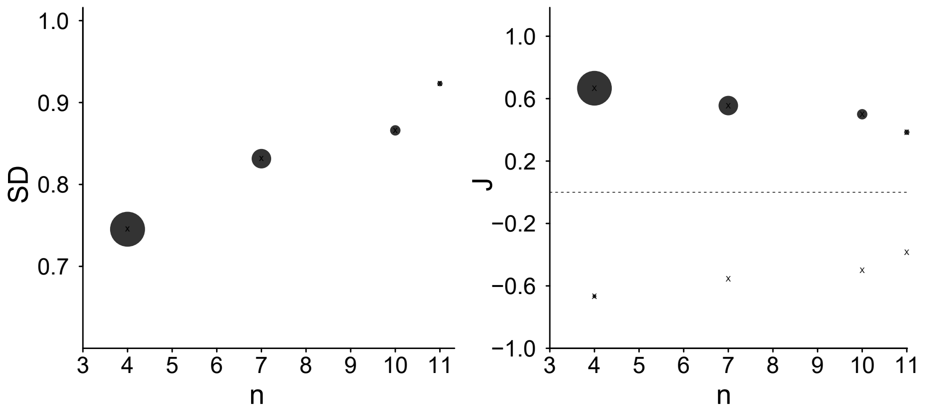


Figure 2. Plot of sample *SD* (left chart) and judgment strength *J* (right chart) as a function of *n*, the size of samples truncated according to a Bayesian stopping rule, assuming truncation when 90% posterior highest density becomes narrower than the threshold interval width $t = .4$. *SD* and *J* are calculated from $p' = \frac{\alpha}{\alpha+\beta}$, the Bayesian estimate of *p("likeable")|n members)*, $SD = \sqrt{p'(1-p'*2}$; $J = 2p'-1$ if $p \geq \frac{1}{2}$; and $J = -2p'+1$ if $p < \frac{1}{2}$.

form priors) multiplied k times with a likelihood ratio of $LR =$ *p(likeable trait added |"likeable" group)/p(likeable trait added |"unlikeable" group)* and $(n\breve{}k)$ times with the reverse ratio $1/LR$. Because k LR updates and $(n\breve{}k)$ $1/LR$ updates cancel each other out, a Bayesian stopping rule is sensitive to the frequency difference of likeable minus unlikeable members, regardless of n. Thus, the same 2:1 ratio of 16 likeable and 8 unlikeable traits in a larger sample affords stronger evidence for the prevailing positivity than 8 likeable and 4 unlikeable traits.

We realize uniform priors as a beta distribution $Beta(\alpha = 1, \beta = 1)$. Updating these priors with k "likeable" and $(n-k)$ "not likeable" traits, the posterior probability density distribution becomes $Beta(\alpha_0 + k, \beta_0 + n - k)$.[2] Sampling is truncated when the highest density of this posterior distribution is condensed within a sufficiently narrow interval. The width of this interval is an expression of how confident and stable the current impression is. Truncating samples this way produces similar (increasing) SD and (decreasing) J functions of n as the Haldane algorithm (Figure 2). Thus, whether the stopping rule is sensitive to $\sqrt{n}$ or not, the simulated self-truncation effect on homogeneity and extremity of group impressions remains largely unchanged. We obtain similar decreasing *SD* and *J* functions in a simulation of sampling from a universe of non-binary likeability scale values. Self-truncated small samples are regularly more homogeneous (consistent, conflict-free) leading to stronger (more polarized) judgments than larger, non-truncated samples.

Empirical Evidence

Encouraged by these simulation results and by converging evidence from earlier simulations and experiments (Prager et al., 2018; Prager & Fiedler, in press), we based our investigation of group impression judgments on the two-stage process model depicted in Figure 3. Parameters of the population sets (particularly the positivity proportions p), from which group traits are sampled, are supposed to exert a direct influence on

[2] In general, across all possible $p(”likeable”)$ levels, $Beta(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} p(”likeable”)^{\alpha-1}(1 - p(”likeable”))^{\beta-1}$ for $0 \leq p(”likeable”) \leq 1$.

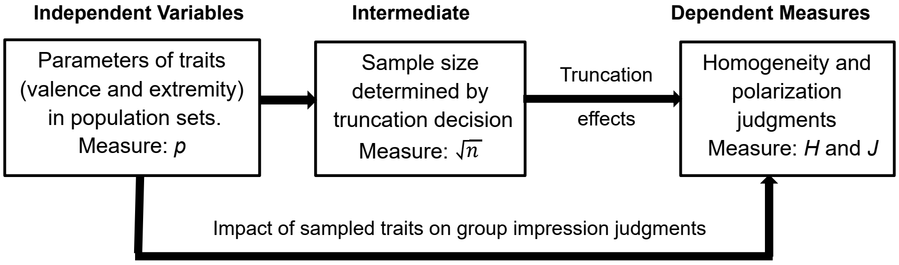

Independent Variables
Intermediate
Dependent Measures
Parameters of traits (valence and extremity) in population sets. Measure: *p*
Sample size determined by truncation decision Measure: $\sqrt{n}$
Truncation effects
Homogeneity and polarization judgments Measure: *H* and *J*
Impact of sampled traits on group impression judgments

Figure 3. Two stage process model of sample-based group impression judgments.

judgments of group homogeneity and polarization (measured by H and J), but also an indirect effect mediated by small sample size ($\sqrt{n}$) resulting from early truncation.

We conducted four experiments to elucidate the antecedent sampling input and the judgmental consequences of sample truncation and to disentangle the direct and the mediate sampling effects on group judgments. As a pre-condition of the main experiments, we ran two pre-studies to check on the suitability of the stimulus materials and task procedures. The purpose of the first pre-study was to demonstrate that sample size n reflects the same stimulus characteristics and produces equivalent judgments effects, regardless of whether it refers to existing groups labeled with common names or to samples without any reference to existing groups. The same canonical relationship of sample size to homogeneity and evaluation strength can be expected with either type of sample. The second pre-study will then establish the applicability of the density model to the trait stimuli used in the following experiments, using a method introduced by Goldstone (1994).

Pre-Study 1: Sample size, Homogeneity, and Evaluation of Existing Social Groups

We selected labels of 28 social groups from Study 3a in Koch et al. (2016), for example “artists”, “car drivers”, “conservatives” or “vegans”. We selected group labels that were supposed to be familiar to most people, but also sufficiently similar in their level of abstraction (see Appendix for the full list). Two subsets of participants were asked to rate either group knowledge or group likeability and homogeneity. Group labels were presented one after another. After seeing a new group label, knowledge ratings were prompted by two questions (on the same screen below the displayed label): “How much do you know about this group in general?” and “How many members of this group do you know?” They responded by clicking on a continuous scale below each question, which was marked at the endpoints, ranging from “nothing” to “very much”, and from “none” to “a great many”. A second subset of participants also completed sequential ratings of one group label after another. This time groups were rated on likeability and homogeneity, prompted by the questions “How much do you like this group?” (continuous scale with endpoints labeled “strong antipathy” and “strong sympathy”) and “How similar are group members to each other?” (endpoints “very different” and “very similar”).

Results and Discussion. We averaged responses to both knowledge questions, which were highly correlated (hierarchical correlation nested in $N = 119$ participants: Mean $r = .77$, $SD = .18$). Individual correlations were positive for 99% of all participants. Likeability (N = 83) and homogeneity (N = 119) ratings for each social group label were averaged across all participants. As expected, group knowledge correlated negatively with rated homogeneity ($r = -.58$, $t(26) = 3.62$, $p = .001$) but positively with likeability ($r = .42$, $t(26) = 2.38$, $p = .025$), reflecting the joint dependency of both major inter-group biases on the amount of group knowledge, operationalized by two highly correlated ratings.

Pre-Study 2: Density of Traits

In the second pre-study, we validated the assumption that the density (i.e., average distance to other stimuli) is higher for positive than for negative traits and for extreme than for moderate traits. This amounts to expecting lower density for traits high than low in diagnosticity.

Methods. The stimulus traits were a set of adjectives that had been scaled for valence in Prager and Fiedler's (in press) experiments on individual impression-formation. We applied the spatial arrangement method (Goldstone, 1994) to assess the distance between the 70 traits of the entire set as the core feature of density. In each of five successive study rounds, participants were asked to position twelve traits (font size 24 pt) in a white square (side length 278 px). They were thoroughly instructed to group fitting words together and to spatially separate non-fitting ones. The white square was initially empty; participants could then drag and drop traits one after another from a grey box below the positioning area by clicking, moving and releasing the mouse. Traits could be relocated later. After all twelve items were placed, a button to continue emerged. Clicking the continue-button cleared the screen and started a new round of twelve traits. For each round the twelve traits were drawn randomly without replacement from the entire pool of traits. Thus, the order of traits was newly randomized for each participant.

The study was conducted first in the context of five other unrelated studies (on serial back-translation, simultaneous encoding of two trends, personal data privacy, and multiple contingencies learning) at Heidelberg University. The age of the ninety-five participants ranged from 18 to 65 years (M = 24.93), with 78 being female, 16 male and one of other gender. Ninety participants were students (16 Psychology students). We intended to use a minimal duration of 20 seconds per round as exclusion criterion, but no participant had to be excluded by this criterion.

The whole procedure was executed by Java software. For each possible pair out of the set of 70 traits, the Euclidean distance between placement positions was averaged across all participants who rated that pair of traits in the same round.

Results. We computed the distance of each trait to the rest of the entire set of 70 by the sum of squared distance values of the individual trait and all other traits. Going beyond Unkelbach et al. (2008), we not only expected negative traits to be more distant from each other than positive ones, but also extreme traits to be more distant from others than moderate ones. In a regression analysis of distance scores, this two-fold expectation should be evident in a strong linear trend (i.e., distance decreasing with trait positivity) along with a quadratic trend (to capture the higher distance of extreme than moderate traits of either valence). Figure 4 corroborates exactly this pattern. Standardized regression coefficients confirmed that distance decreased with increasing positivity (linear $\beta = -.64$, $t(68) = 6.84$, $p < .001$) and with increasing extremity (quadratic $\beta = .21$, $t(68) = 2.28$, $p = .026$).

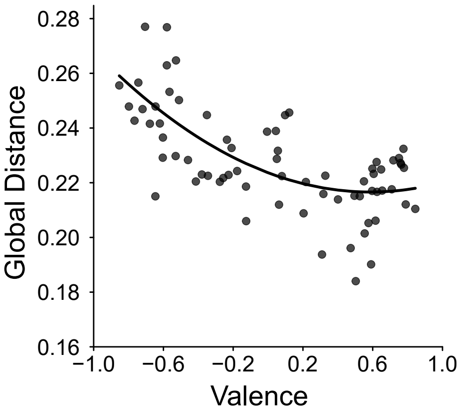


Figure 4. Average distance between traits and the remaining set (i.e. "global distance") plotted as a function of trait valence norms. Each point represents one trait term.

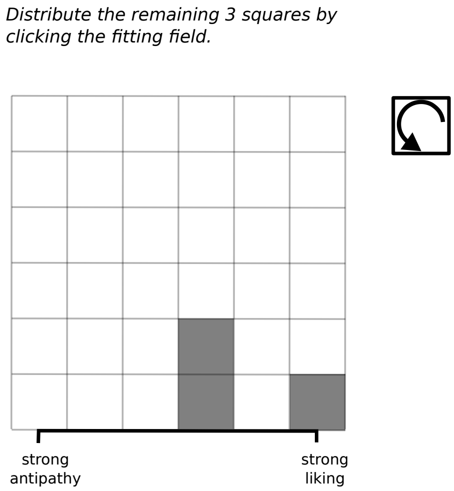
Experiment 1

Having approved the stimulus materials and the comparability of real groups and mere trait samples, we can now move on to Experiment 1. Its primary purpose was to substantiate the assumptions of the two-stage process model in Figure 3. In essence, we wanted to demonstrate that similar self-truncation effects as in former research by Prager et al. (2018) on impressions of individual targets can be found in a group impression formation task of less entitativity, when each sampled trait refers to a different member of a group. Moreover, because inter-group research is concerned with both homogeneity and impression strength, Experiment 1 introduces homogeneity measures. We assessed both mean impressions and perceived homogeneity using two operationally independent measures: rating-scales and the distribution-builder method (Sharpe et al., 2000). We expected judgments of group homogeneity and polarization (on the likeability scale) to be jointly determined by a direct influence of the traits sampled from the population set and an indirect influence mediated by the sample size resulting from self-truncation (see Figure 3).

To substantiate the viability of mere trait samples for the study of group impressions, we compared one condition with meaningful labels of existing social groups to another condition with meaningless labels (like "group A") attached to mere samples of traits. Based on materials constructed in previous research (Prager & Fiedler, 2021) on individual impression formation, we relied on population sets of traits that represented different levels of diagnosticity (valence and extremity). All four experiments involved random sampling of traits from these distinct population sets.

Participants and design. One-hundred and thirty-four participants were recruited from a participant pool at Heidelberg University. Participants' age ranged from 17 to 77 ($M = 24.88$); 107 participants were female. One hundred and twenty-six participants were students, of which 36 were students of psychology. Fifteen participants who had invariantly sampled one or all 16 traits on every trial were excluded. Of the remaining participants, 62 were randomly assigned to the meaningful groups condition and 57 to the meaningless-labels condition. In addition to this between-participants factor, the positivity proportion p of the population set from which trait samples were drawn was manipulated within participants as a repeated-measures factor.

The experiment was second in a sequence of four unrelated experiments (on directed forgetting, speed-accuracy trade-off, and simultaneous encoding of two trends) in a



Distribute the remaining 3 squares by clicking the fitting field.

Figure 5. The distribution builder: A participant has already placed three squares on a moderately and an extremely positive likeability-position. Three further squares must still be set.

one-hour session. Participation was compensated either by payment (8€) or by course credit.

Materials. From the entire list of 70 adjective traits (see pre-study), we extracted four potentially overlapping population sets of 30 randomly ordered stimulus traits, from which the stimulus samples were drawn. The four sets had positivity proportions of p = .20, .33, .67, .80[3], that is two sets were predominantly positive, two negative in valence and orthogonally, two sets were of moderate and two of extreme valence. On each trial, a target group was described by a random sample of traits drawn from one population set. Each participant completed 28 trials in random order, seven drawn from each population set.

Procedure. After providing basic demographic data, participants received instructions saying that each trial started with a group label displayed on top of the screen (21pt bold face). By pressing the space bar, participants could invoke a new trait (displayed in font size 20pt in the top center region of the screen). At each point between $n = 1$ and $n = 16$, they could decide to either solicit another trait (by pressing the space bar again) or to truncate the sample using the Enter key and proceed to the judgments. All traits remained on screen as long as sampling continued. They were listed vertically; when sample size exceeded $n = 8$, additional traits were displayed in a second column. The most recent trait was highlighted by enhanced contrast.

Immediately after the truncation decision, participants provided their group judgments. Using a modified version of the Sharpe et al. (2000) "distribution builder", they were asked to construct a distribution curve for the group's likeability, by clicking the cells of six degree-of-liking columns of a 6 x 6 grid (illustrated in Figure 5). Clicking a column results in placement of a grey square in the respective column. Note that the distribution builder offers a measure of both the mean and dispersion of the likeability impression on the target group.

Upon completion of the distribution builder task, participants continued to a new screen on which three horizontal graphical rating scales with marked endpoints appeared

[3] Mean (SD) of the four population sets were -.28 (.39), -.19 (.41), .18 (.42), .30 (.42). Selection of traits aimed at keeping all parameters identical or symmetrical between sets.

Table 1. Regression Analyses Testing for Consistency and Sensitivity of Impression Judgments Assessed by the Distribution Builder and Rating Scales

	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Tests for consistency				
$r(M_{DB}, L_{Rating})$	.90 (.13); 99%	.89 (.09); 100%	.92 (.06); 100%	.92 (.10); 100%
$r(SD_{DB}, H_{Rating})$	-.38 (.28); 87%	-.43 (.26); 92%	-.33 (.36); 79%	-.46 (.26); 96%
Test for sensitivity to sampling input				
$r(L_{Rating}, M_{Trait\ valence})$	.77 (.21); 98%	.79 (.19);98%	.77 (.16); 100%	.84 (.11); 100%
$r(M_{DB}, M_{Trait\ valence})$	.74 (.20); 99%	.76 (.19); 98%	.75 (.15); 100%	.81 (.12); 100%
$r(H_{Rating}, SD_{Trait\ valence})$	.30 (.21); 91%	.32 (.22); 96%	.27 (.31); 77%	.38 (.23); 94%
$r(SD_{DB}, SD_{Trait\ valence})$	-.30 (.23); 90%	-.36 (.26); 88%	-.31 (.28); 84%	-.38 (.24); 92%

Note. The upper part of the table provides correlations (r) between distribution builder (DB) measures and ratings of likeability (L) and homogeneity (H). The bottom part contains means (M) and standard deviations (SD) of individual judges' correlations between measures of L and H and average trait valence norms across all 28 samples. Percentages indicate consensus proportions (i.e. the proportion of participants' individual values sharing the sign of the average value) of individual judges' correlations sharing the same sign as the average correlation.

one after another. They provided their ratings by clicking on the appropriate horizontal scale position. Scales referred to likeability ("How much would you like that group" with endpoints "highly unlikeable" and "highly likeable"), homogeneity ("How similar are the group members in their likeability?", endpoints "very different" to "very similar"), and confidence in both ratings ("How confident are you in these judgments?" ranging from "very uncertain" to "very sure"). During the entire judgment phase, the current group label remained visible in the central top position on the screen. When all ratings were completed, participants could start sampling traits of a new group by clicking a button[4].

Results

Quality of judgment data. Table 1 provides an overview of the quality and reliability of the judgment data from all four experiments. Consistency checks testify to the high quality of judgments and to participants' motivation. We extracted means and standard deviations of likeability ratings L from the distribution-builder responses to the six likeability scale values: (-1, -.6, -.2, .2, .6, 1), weighted by the number of squares that had been placed in the respective column. Individual participants' correlations between the average likeability positions of the distribution builder and the likeability ratings on the graphical scale ranged from $r = .89$ to .92.

For an index of sensitivity to sampled input traits, we correlated ratings and distribution builder scores with the average valence norms of the sampled traits underlying the likeability judgments.[5] The correlations in the bottom of Table 1 corroborate that judgments were highly sensitive to the direct influence of the sampled traits' valence norms. Likeability measures correlated strongly with average valence scale values of the sampled traits (average r ranged from .74 to .84 across experiments). Consensus rates across participants were close to 100% (see Table 1).

Both measures of homogeneity converged only moderately: The standard deviation of distribution-builder positions (i.e. inverse homogeneity) correlated with homogeneity assessed by the rating scales in the range of $r = -.33$ to -.46 between the four experiments. The homogeneity measures correlated only moderately with samples' valence

[4] After finishing the experimental task, participants worked on knowledge ratings of pre-study 1a.

[5] As in the preceding quality checks, these results refer to likeability judgments L as they were, as distinguished from the deviation scores J used to assess judgment strength.

Table 2. Means (SD) of Individually Calculated Standardized Regression Weights (β) Using the Linear and Quadratic Trend of Population Set Valence (Proportion Positive) as Predictors.

	Predictors	
Criterion	p (linear)	p (quadratic)
Sample size $\sqrt{n}$	$\beta = .26$ (1.15) $t(118) = 2.50$, $p = .014$ 62%	$\beta = -.24$ (1.13) $t(118) = 2.28$, $p = .024$ 62%
Homogeneity H	$\beta = -.48$ (1.09) $t(118) = 4.82$, $p < .001$ 66%	$\beta = .57$ (1.06) $t(118) = 5.80$, $p < .001$ 71%
Impression strength J	$\beta = -1.02$ (1.08) $t(118) = 10.30$, $p < .001$ 82%	$\beta = .83$ (1.05) $t(118) = 8.63$, $p < .001$ 80%

norm standard deviations (i.e. inverse homogeneity; average r from .27 to .38).

Valence and extremity as antecedents of truncation and direct sampling effects. The four population sets from which samples were drawn allowed for an orthogonal test of the impact of two valence-measures of diagnosticity. When predicting $\sqrt{n}$, H and J from the proportion p of positive traits in the population sets (p = .20, .33, .67, .80), the influence of negative versus positive valence should be manifested in a linear trend whereas the influence of extreme versus moderate valence should be evident in a quadratic trend.

Consistent with the expectation that more diagnostic traits facilitate earlier truncation, regression analyses of $\sqrt{n}$ showed that (the square root of) sample size tended to be smaller for negative (linear β: $M = .26$, $SD = 1.15$) and for extreme population sets (quadratic β: $M = -.24$, $SD = 1.13$) than for positive and moderate population sets. Likewise, the analysis of judgment strength confirmed the expectation of a direct influence. Negative and extreme sets evoked stronger impressions than positive and moderate ones (linear β: $M = -1.02$, $SD = 1.08$; quadratic β: $M = .83$, $SD = 1.05$). Moreover, in the analysis of perceived homogeneity H, samples/groups described by negative and extreme compared to positive and moderate population sets appeared more homogenous (linear β: $M = -.48$, $SD = 1.09$; quadratic β: $M = .57$, $SD = 1.06$; see Table 2).

Truncation effects on H and J. Turning to the self-truncation effects proper (i.e., higher homogeneity and group polarization in small than in large groups), we calculated within each individual participant (across trials) the correlations of samples size $\sqrt{n}$ with impression strength J and homogeneity H. We averaged all measures (distribution builder and ratings) of impression strength J and both (z-standardized) values of homogeneity H. Recall that impression strength J is a deviation score with a positive sign when likeability L deviates from the scale midpoint in the correct direction (i.e., $L >$ midpoint for samples drawn from a positive set or $L <$ midpoint for samples drawn from a negative set; J is negative for deviations pointing in the incorrect direction). We related the square root $\sqrt{n}$ of sample size to homogeneity and strength of impression judgments to capture the expected non-linear function.

The scatter diagrams in Figure 6 reflect the typical impact of self-truncated sampling. Table 3 indicates the corresponding mean sample sizes (and standard deviations) of $\sqrt{n}$, homogeneity scores H, impression strength scores J, and average individual correlations $r_{\sqrt{n},J}$, $r_{\sqrt{n},H}$, and $r_{J,H}$, (across 28 judgments) along with the standard deviations (between participants) in parentheses. Results in the left column are pooled over all participants; separate results for meaningful groups and meaningless samples

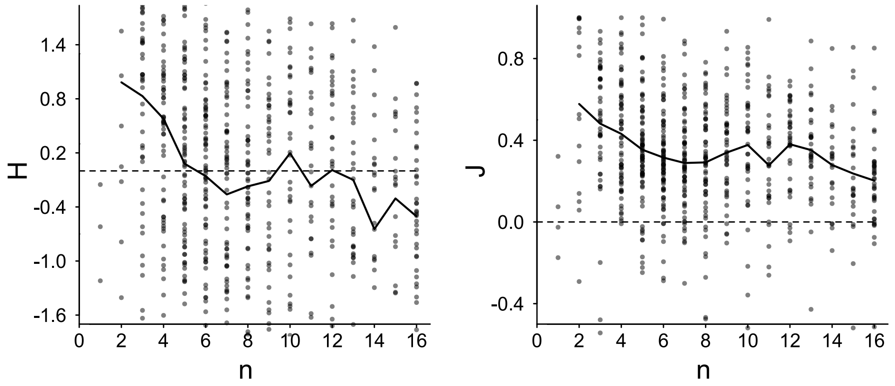


Figure 6. Perceived homogeneity H (left) and impression strength J (right) plotted as function of self-truncated sample size n. Grey dots indicate individual averages for each sample size, the black line represents global averages split by sample size.

Table 3. Means (SD) and % Consensus in Sign Across Participants

Variable	Pooled for both conditions	Meaningful social group labels	Meaningless labels
n	7.68 (3.74)	7.69 (3.90)	7.67 (3.60)
$\sqrt{n}$	2.67 (.66)	2.67 (.68)	2.67 (.64)
H	0 (.76)	-.10 (.79)	.11 (.73)
J	.28 (.11)	.25 (.12)	.32 (.09)
$r_{\sqrt{n},H}$	-.19 (.25); 74%	-.17 (.25); 73%	-.22 (.25); 81%
$r_{\sqrt{n},J}$	-.19 (.23); 82%	-.16 (.23); 74%	-.22 (.20); 88%
$r_{,H,J}$	.44 (.25); 93%	.41 (.26); 92%	.47 (.24); 95%

are given in the middle and right column, respectively.

Several sensible findings deserve to be emphasized. First, a strong overall tendency towards positive J scores shows that participants were highly sensitive to actually existing trends in the population sets from which trait samples were drawn. Yet, secondly, there was considerable variation across judgments in $\sqrt{n}$, J and H to allow for marked correlations. The last row of Table 2 shows that impression strength J and homogeneity H correlated strongly and positively over trials. Of central importance, third, both self-truncation effects were evident from clearly negative correlations, $r_{\sqrt{n},H}$ and $r_{\sqrt{n},J}$, reflecting both small-group homogeneity and small-group polarization. Mean correlations of roughly one standard deviation below zero reflected a strong and regular effect that generalized across a large majority of participants: mean $r_{\sqrt{n},H} = -.19$, $SD = .25$ (consensus in sign: 76%); mean $r_{\sqrt{n},J} = -.19$, $SD = .22$ (consensus 82%).

Knowledge: Effects of meaningful group labels. A comparison of the middle and right column of Table 3 shows that both experimental conditions yielded highly similar results. Whether trait samples referred to naturally existing groups with a familiar name or to samples with meaningless names, the same fundamental self-truncation effects emerged. Small sample size triggered high perceived homogeneity and strong impressions. A small $\sqrt{n}$ in a meaningless sample yielded similar small-group homogeneity as impoverished knowledge about existing out-groups. A contrast in $r_{\sqrt{n},H}$ between conditions was not significant, $t(117) = 1.06$, $d = .19$, $p = .291$.

Table 3 also shows that group labels did not affect sample size ($t(117) = .04$, $p = .968$) and exerted little influence on homogeneity ($t(117) = 1.51$, $p = .135$); samples labelled with existing social group names appeared slightly less homogeneous than groups without meaningful labels. Still, judgments of existing social groups were less polarized, that is, groups labelled with existing names were judged less strongly ($t(117) = 3.64$, $p < .001$).

Discussion

Regular and consistent results corroborate the reliability of various measures of group impression judgments obtained in Experiment 1, reflecting participants' motivation and commitment. The average pretest valence scores of the sampled traits afforded a powerful predictor of likeability judgments, and so did dispersion of sample valence norms predict homogeneity reports: Judgments were systematically driven by the sampling contents.

The results obtained with the "distribution builder" (Sharpe et al., 2000) strongly converged with the continuous rating measures of group impression judgments, suggesting that the distribution builder offers a valuable instrument for group-judgment research.

Experiment 1 suggested that the same functional rules that describe impression judgments of individuals from trait samples (Prager et al., 2018) also apply to groups as impression targets. We consolidated the sample-truncation effect obtained for individual impression judgments. Impressions were stronger and more homogenous for smaller than for larger samples, analogous to the well-known polarization and homogenization of (small) out-group samples, relative to (large) in-group samples (Brewer, 1993; Linville & Jones, 1980; Quattrone & Jones, 1980).

Comparisons of judgments of existing groups with meaningful labels to mere trait samples of unknown "groups" with meaningless labels did not reflect a systematic influence of unequal prior knowledge. Sample-truncation effects (small-sample homogeneity and polarization) were about equally strong in both experimental conditions. Yet, meaningful group labels led to less extreme, more cautious likeability judgments than samples without meaningful labels, presumably because prior knowledge serves to dampen the impact of on-line sampled trait information.

The experimental design included two valence-related measures of diagnosticity: population set valence and extremity. Despite the reduced entitativity of groups compared to individuals, which might dilute systematic influences on sample truncation and impressions, group judgments nevertheless showed the same typical tendency as individual judgments towards earlier truncation and stronger judgments when negative and extreme traits rendered information more diagnostic than positive and moderate traits. Consistent with our theoretical reasoning, high diagnosticity of the sampled traits enhanced the impact of sample size and truncation effects.

Experiment 2

So far, Experiment 1 and two pre-studies provided convergent evidence for both antecedent and consequent conditions of self-truncation. On one hand, because negative and extreme traits are more diagnostic than positive and moderate traits, valence and extremity afforded critical predictors or antecedent conditions of self-truncation. On the other hand, the small sample sizes resulting from early truncation had two regular consequences; small sample size $\sqrt{n}$ led to high homogeneity H and high impression strength J. Moreover, as anticipated by the structure depicted in Figure 3, our findings also reflect a direct influence of sampled traits on group judgments, independent of the mediational truncation effect, as manifested in stronger trait valence and extremity effects on H and J than on self-truncated sample size $\sqrt{n}$.

Nevertheless, the stable and coherent pattern of Experiment 1 results raises some open questions concerning the relationship of the two major phenomena of inter-group judgments, out-group homogeneity and out-group polarization. One problem refers to the independence of H and J measures. To the extent that the variance of extreme population sets (at positivity proportions of $p = .2$ and $p = .8$) is restricted compared to the variance of moderate population sets (at $p = .33$ and $p = .67$), maybe in particular for negative traits, the homogeneity effect may reflect an artificial consequence of the ecological population from which traits are sampled.

To disentangle perceived homogeneity, H, conceived as a dependent measure or a judgmental consequence, from the variability of traits in the population sets, in Experiment 2 we deliberately manipulated variability orthogonally to the valence of the population sets. Thus, at each level of positivity, we constructed two parallel population sets of traits, one with traits of high variability in their likeability values, and one with traits of low variability. If the perceived homogeneity continues to be higher for small than for large groups in such an orthogonal design, this can no longer reflect the restricted variance of extreme population sets.

More generally, the orthogonal design allows us to understand the direct as well as the indirect (truncation-mediated) influence of diagnosticity (valence and extremity) on group judgments. The density model implies that negative and extreme traits should trigger earlier truncation, yielding smaller sample sizes, and higher H and J judgments than positive and moderate traits, for higher distance between the former traits increases diagnosticity (i.e., lesser overlap of added traits with preceding traits). Conversely, there is no reason to expect that the other factor of the orthogonal design, high versus low variability of trait likeability in population sets, leads to faster truncation or stronger H and J judgments, simply because density relies on multidimensional trait distance, rather than on uni-dimensional likeability "distance". The design guarantees that high and low likeability variability of positive and negative population sets is symmetrical. Our theoretical approach implies a persistent influence of diagnosticity (valence x extremity) even when trait variability is controlled for.

Methods

Participants and design. Ninety-four participants were recruited for the second experiment via the Psychology subject pool at Heidelberg University. The experiment was conducted as an online study, which was controlled by php and javascript. Participants' age was 18 to 59 years (23.56 years on average), of which 72 identified themselves as female, 20 as male and 2 as other. Eighty-seven students (22 psychology students) participated. We also excluded data from two participants whose total participation time was shorter than 7 minutes (median around 13 minutes), which did not allow for careful task execution. Another 11 participants needed to be excluded from data analysis because their sample truncation strategies resulted in a constant sample size of either $n = 1$ (minimum) or $n = 16$ (maximum), resulting in 82 analyzed data sets.

Both design factors, the positivity proportion p of traits in different population sets and trait variability, varied within participants, across 20 impression judgment trials, as explained in the Materials section below.

Materials. When selecting items from the pool of traits, variability and valence are hard to separate. Trait populations of extreme valence tend to be more homogenous in valence than moderate ones, because upper and lower boundaries restrict distributions in extreme ranges. To construct population sets of orthogonal valence and homogeneity, we applied an iterative procedure: As the maximum observable sample size was set to 12 traits, we set the population set size to 12 accordingly. To manipulate valence, we set five levels of p, involving 2, 4, 6, 8, and 10 positive out of 12 traits. For each of these five valence levels, we formed parallel population sets: two of (identical) high and two of low within-set variance each using an iterative algorithm (repeated random draws with replacement from the total set of 70 traits) until the resulting population sets approximated the desired characteristics (symmetry in valence mean, approximate equivalence in density and frequency). The iterative procedure made sure that not only the entire set but also sub-set parameters of each population set remained stable with regard to the parallel variability levels, valence, and control measures. By generating two sets per parameter combination (i.e. 5 p-levels x 2 variability levels), 20 population sets were formed in total. A side benefit resulted from this procedure: Valence now varied at a slightly higher level of resolution than before; the 20 population sets now covered 5 steps of negative versus positive valence (i.e., p varying from 2/12 to 10/12).

Procedure. The entire procedure was largely the same as in Experiment 1. Each participant provided judgments of 20 self-truncated trait samples representing all 20 population sets, with meaningless group labels. Set order within the experiment and the order of traits within each set were shuffled (randomized) for every participant.

Results

Valence and extremity as antecedents of truncation and direct sampling effects. To replicate the basic results from Experiment 1, we ran regression analyses with criteria $\sqrt{n}$ (truncation) and then of H and J on the valence and extremity predictors (diagnosticity), operationalized by linear and quadratic trend over the p parameter in addition to the predictor within-set variability. Note that all three predictors, valence (linear p), extremity (quadratic p), and trait variability, are orthogonal by design, thus making interaction terms obsolete. Table 4 and Figure 7 summarize the results. Regression parameters for the linear and quadratic trend of valence p consolidate findings of Experiment 1. Samples tend to be truncated earlier (criterion $\sqrt{n}$) the more negative and extreme the population set is in valence. Accordingly, samples from predominantly negative and extreme in comparison to positive and moderate sets are judged more strongly (J) and more homogenous (H).

Within-set variability. The novel variable of this experiment, within-set variability at different p levels, hardly explained any variance when predicting $\sqrt{n}$ and J, beyond valence (linear p effect) and extremity (quadratic effect). Not surprisingly, the variability predictor did receive a significant regression weight for the prediction of perceived homogeneity H, which is evident from the divergence of curves in the center chart of Figure 7. Nevertheless, H judgments were mainly determined by the diagnosticity of the sampled traits; valence and extremity together accounted for more variance in perceived homogeneity than population set variability. Note also that the curves for high and low variability diverge mainly for positive impressions, when high variability entails the diagnostic impact of a few negative traits in a set.

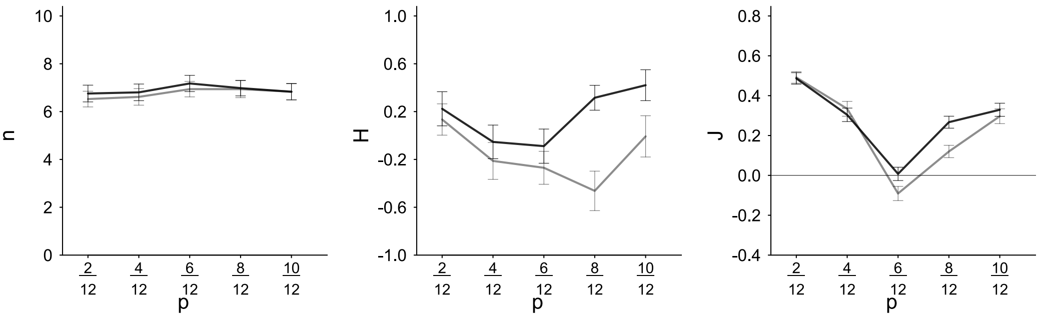


Figure 7. Sampling and judgment parameters (sample size, Homogeneity H, impression strength J) plotted by population sets. Population set proportion of positive traits p is plotted on the abscissa. Lines connect the means of participants' individual values; error bars represent specific standard errors. Light grey (dark grey) curves refer to high-variance (low-variance) population sets.

Table 4. Standardized Mean (SD) β Parameters of Multiple Hierarchical Regression with Predictors Population Set Variability and Valence (Linear and Quadratic Trend)).

	Predictors		
Criterion	p (linear)	p (quadratic)	Variability
Sample size $\sqrt{n}$	$\beta = .41$ (1.21) $t(81) = 3.06$, $p = .003$ 60%	$\beta = -.34$ (1.19) $t(81) = 2.61$, $p = .011$ 61%	$\beta = .04$ (.20) $t(81) = 1.72$, $p = .089$ 59%
Perceived homogeneity H	$\beta = -.61$ (1.22) $t(81) = 4.50$, $p < .001$ 68%	$\beta = .65$ (1.20) $t(81) = 4.92$, $p < .001$ 67%	$\beta = .11$ (.23) $t(81) = 4.17$, $p < .001$ 70%
Impression strength J	$\beta = -1.93$ (1.00) $t(81) = 17.42$, $p < .001$ 98%	$\beta = 1.82$ (.98) $t(81) = 16.81$, $p < .001$ 99%	$\beta = .06$ (.20) $t(81) = 2.90$, $p = .005$ 67%

Note. Each row reports results of one multiple regression analysis, respective criteria are indicated by the left column. Percentages refer to the proportion of individual parameters of the same sign as the aggregate coefficient.

Experiment 3

Experiment 3 offers an immediate test to corroborate the density account of truncation and group impression judgments. If the average distance of traits to all other traits is the crucial stimulus property that drives the impact of sampling and truncation on group judgments, then the distance norms should explain for a substantial part of variance in regression analyses of $\sqrt{n}$, H, and J, beyond the valence and extremity predictors. For a test of this notion, in Experiment 3 we replaced the manipulation of population set variability by a manipulation of average trait density within population sets (and thus samples), which was again orthogonal to the manipulation of the p proportions underlying the manipulation of valence and extremity. If trait diagnosticity is sensitive to the distance parameter of stimulus traits, capturing the diversity versus overlap of traits in multidimensional space, then the average distance norm of trait samples should contribute substantially to predicting $\sqrt{n}$, H, and J.

Methods

Participants and design. Ninety-eight participants were recruited via the Psychology subject pool at Heidelberg University. The experiment was the second in a block of three unrelated studies (on tradeoff-decisions in an information-purchasing paradigm, and on environmentally sustainable behavior) at a Psychology lab at Heidelberg University. Participants were between 17 and 77 years old (mean = 26.11), 75 were female. Ninety students (14 Psychology) participated. Five participants, who consistently truncated sampling after one item or did never truncate before automatic stopping (at $n = 16$) were excluded from the analyses; 93 data sets were analyzed. As in Experiment 2, both design factors, the positivity proportions p and the trait distance norms at each level of p varied within participants, across 34 judgment trials.

Materials. Analogous to Experiment 2, the aim was to generate population sets of orthogonal within-set distance, valence (linear p), and extremity (quadratic p). Although positive (negative) valence and high (low) density are naturally related in the ecology (see Unkelbach et al., 2008), we relied on a similar iterative procedure as in Experiment 2 to accomplish an orthogonal manipulation. Setting the maximum observable sample size to 16, we defined 17 levels of p, ranging from 0 positive traits up to 16 out of 16 positive traits. For each of these 17 valence levels, we formed two parallel population sets: one of high and one of low average within-set density. The population sets were selected by an iterative sampling algorithm, where possible population sets were repeatedly drawn (with replacement) from all available traits, but only those with the best-fitting characteristics were kept. Those target characteristics were symmetry in valence, identical values in word frequency, but also within-set stability (i.e. that sub-sets of the total set have similar density-values as the total set).

Procedure. Each participant engaged in self-truncated sampling and judgments on 34 unlabeled groups corresponding to the 34 population sets. Set order within the experiment and the order of traits within the sets were both shuffled randomly for every participant.

Results

Valence, extremity and within-set density effects. Again, valence and extremity strongly predicted all three criteria $\sqrt{n}$, H, and J in a series of regression analyses, thus

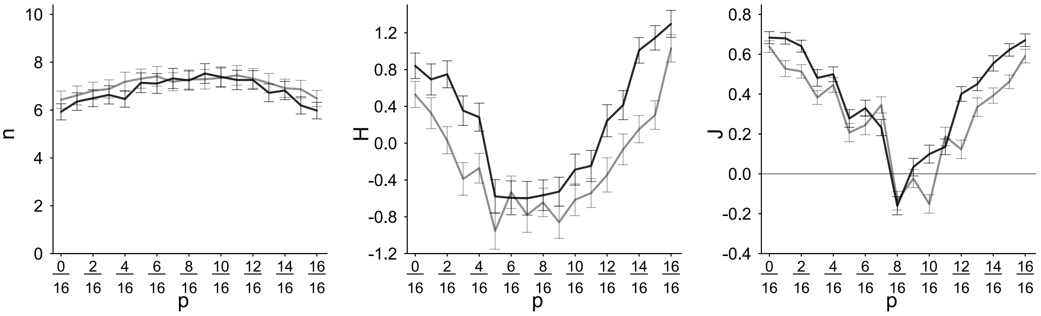


Figure 8. Sampling and judgment parameters (sample size, Homogeneity H, impression strength J) plotted by population sets. Population set valence (p) is plotted on the abscissa. Lines connect the means of participants' individual values; error bars represent specific standard errors. Light grey indicates diverse and dark grey dense population sets.

Table 5. Standardized Mean (SD) β Parameters of Multiple Hierarchical Regression with Predictors Population Set Valence (Linear and Quadratic Trend)) and Average Within- Population Set Density.

	Predictors		
Criterion	p (linear)	p (quadratic)	Within-set density
Sample size $\sqrt{n}$	$\beta = .82$ (.80) $t(92) = 9.93$, $p < .001$ 82%	$\beta = -.79$ (.80) $t(92) = 9.56$, $p < .001$ 84%	$\beta = -.07$ (.18) $t(92) = 3.54$, $p = .001$ 66%
Homogeneity H	$\beta = -1.17$ (.86) $t(92) = 13.12$, $p < .001$ 87%	$\beta = 1.32$ (.82) $t(92) = 15.49$, $p < .001$ 95%	$\beta = .14$ (.17) $t(92) = 7.64$, $p < .001$ 78%
Impression strength J	$\beta = -1.94$ (.52) $t(92) = 36.29$, $p < .001$ 100%	$\beta = 1.74$ (.51) $t(92) = 33.03$, $p < .001$ 99%	$\beta = .10$ (.13) $t(92) = 7.33$, $p < .001$ 80%

Note. Each row reports results of one multiple regression analysis, respective criteria are indicated by the left column. Percentages refer to the proportion of individual parameters of the same sign as the aggregate coefficient.

replicating the same robust pattern as in previous experiments. However, including the average distance score of sampled traits as a third predictor, along with the linear (valence) and quadratic (extremity) p-contrast, created a marked increment in predictive validity. As portrayed in Figure 8 and in the descriptive statistics in Table 5, truncation ($\sqrt{n}$) as well as group impressions (H and J) were not only strongly related to valence and extremity but also to the distance predictor.

Low within-sample distance (i.e. high density) lead to earlier truncation than high within-sample distance (low density), as evident from a mean regression weight of $\beta = -.07$ (.18), $t(92) = 3.54$, $p = .001$ and a consensus rate of 66% of participants with a negative individual β. Because group judgments were not only affected indirectly by the truncation effect but also directly reflected the properties of sampled traits (see Figure 8), distance received a stronger regression weight in predictions of H, mean $\beta = .14$ ($SD = .17$), $t(92) = 7.64$, $p < .001$, 78% consensus, and J, $\beta = .10$ ($SD = .13$), $t(92) = 7.33$, $p < .001$, 80% consensus. Both perceived homogeneity (H) and impression polarization increased with decreasing trait distance.

Thus, consistent with the density-model analysis, the average distance of sampled traits allowed us to predict the truncation measure $\sqrt{n}$ and the greatest part of the group-impression measures J and H, whereas within-set variability (operationalized as variance between traits) did not contribute much beyond the diagnosticity of traits in the previous experiment. After all, diagnosticity (i.e., negative valence and extremity) is a property of traits, independent of the variance of the set of different traits included in a set.

Experiment 4

To round up the insights gained from our density-model analysis of trait sampling and group impression formation, we report a final experiment that included a correlate of density that is presumably irrelevant to trait diagnosticity, namely, the linguistic frequency of trait terms. Previous research has shown that positive high density (low-distance) words occur more frequently in a large text corpus than negative low density (high-distance) words. However, despite this correlation, there is no reason to assume that trait diagnosticity depends on the infrequency of words in the lexicon, as distinguished from their distance in mental representations. It is easy to find synonyms of the same trait (at a given level of diagnosticity) denoted by frequent and infrequent words (example: loyal and trustworthy), and it is easy to think of equally frequent words of clearly different diagnosticity (example: detail-orientated and manipulative). Our theoretical approach therefore predicts that word frequency counts should hardly contribute predictions of $\sqrt{n}$, H, and J from trait samples.

Methods

Participants and design. Fifty-six participants were recruited via the Psychology subject pool of Heidelberg University. The experiment was the first in a 55-minutes session of four unrelated studies (on speed-accuracy tradeoff in choices, detection of statistical suppression, and probability prediction from sampling). Participants' age ranged from 18 to 51 with 22.88 years on average. Thirty-five identified themselves as female, 20 as male and one other. Twelve out of a total of 55 students were students of psychology. Three participants who invariantly sampled one or the maximum of all 16 traits were excluded, leaving 53 data sets for the analysis. In a complete repeated-measures design, each participant received in total 32 samples from four population sets (consisting of 30 traits each), representing orthogonal combinations of positive versus negative valence (corresponding to p = .27 and p = .73) and high versus low word frequency. Note that given two p levels we could not distinguish extremity from valence. Although both factors were related to density, only valence should affect the diagnosticity of traits (conceived as distance in a mental representation), whereas word frequency should not affect sample truncation and group impression judgments.

Materials. We assessed word frequency of the entire pool of 70 traits in large corpora of spoken and written language (Leibniz Institut Für Deutsche Sprache, 2014, 2017), which offered an approximate measure of word frequency. Four population sets were constructed by an iterative algorithm; sets of 30 traits were randomly drawn from the entire pool of 70 traits in a repetitive loop, until frequency and valence were orthogonally distributed, whereas density values were equalized across population sets. Positive population sets had a proportion of $p = .73$, as compared to $p = .27$ for negative sets. Positive and negative sets were selected to be symmetrical also in their valence mean and approximately equal in variance.

Table 6. Means (SD) of Individually Calculated Regression Weights (β) of Three Diagnosticity Contrasts

		Predictors	
Criterion	C1: Valence	C2: Word frequency	C3: Interaction
Sample size $\sqrt{n}$	$\beta = .03$ (.16) $t(52) = 1.28$, $p = .208$ 58%	$\beta = -.04$ (.19) $t(52) = 1.51$, $p = .137$ 60%	$\beta = -.01$ (.17) $t(52) = .40$, $p = .694$ 55%
Homogeneity H	$\beta = .10$ (.22) $t(52) = 3.34$, $p = .002$ 66%	$\beta = .01$ (.19) $t(52) = .43$, $p = .667$ 53%	$\beta = .02$ (.20) $t(52) = .75$, $p = .459$ 51%
Impression strength J	$\beta = -.26$ (.30) $t(52) = 6.34$, $p < .001$ 81%	$\beta = .01$ (.18) $t(52) = .45$, $p = .652$ 45%	$\beta = -.02$ (.17) $t(52) = .75$, $p = .459$ 49%

Note. Each row reports results of one multiple regression analysis, respective criteria are indicated by the left column. Percentages refer to the proportion of individual parameters of the same sign as the aggregate coefficient.

Results and Discussion

Predicting truncation and judgment effects from valence and word frequency. We included three predictors for hierarchical regression analyses for sample size ($\sqrt{n}$) homogeneity (H) and impression strength (J); trait valence, frequency, and the interaction thereof. The interaction term contrasted the population sets negative – infrequent and positive – frequent (naturally occurring combinations) against positive – infrequent and negative – frequent (reversed to the typical relation in language). We averaged distribution builder and rating scores of H and J, as in all previous analyses.

As evident from the last two columns of Table 6, neither word frequency per se nor its interaction with valence was ever a significant predictor, corroborating the notion that the frequency of words in the lexicon is largely detached from the diagnosticity of its referent trait. Only the valence of sampled traits was again related systematically to H and J, although in this particular study we did not support the typical impact of negative valence on earlier truncation. In the absence of a final explanation for this unexpected failure to replicate the dependency of truncation on valence, we tend to attribute this singular abnormal finding to the fact that very large set sizes in Experiment 4 allowed for a very high sampling error. Because of the resulting confusion between trait samples from positive and negative population sets, the enhanced diagnosticity of negative valence only fostered higher H and J judgments, but failed to affect the truncation decision.

As in all previous experiments, truncation bore systematic relations to subsequent judgments. Smaller samples solicited more homogenous group impressions (mean $r_{\sqrt{n},H} = .-.30$, $SD = .27$, consensus in sign: 87%) and stronger impressions than larger samples (mean $r_{\sqrt{n},J)} = -.24$, $SD = .22$, consensus in sign: 87%). Moreover, perceived homogeneity and impression strength were substantially correlated (mean $r_{J,H} = .39$, $SD = .23$, consensus in sign: 94%).

General Discussion

In sum, all simulation results and empirical findings obtained in the present research converge to a robust but nevertheless refined pattern that allows us to draw distinct conclusions. Our research highlights that sample-based impression judgments follow the same set of distinct rules regardless of whether the sampled traits characterize one individual target or whether each trait belongs to a different member of a group. In either case, impression judgments are highly sensitive not only to the parametric influence of the population sets from which the stimulus samples are drawn, but also to the specific traits of individual samples. The final impression judgments were highly predictable from the normative properties of the randomly sampled traits (i.e., from their semantic and pragmatic scale values) that were determined in careful pilot testing.

Notably, the resulting impression judgments systematically deviated from a simple averaging rule; that is, the group's likeability was not a simple average of all sampled traits' likeability scale values (Anderson, 1965). Rather, trait diagnosticity strongly moderated the impact of a newly added trait on the sequential impression updating process. Highly diagnostic traits exerted a stronger impact on integral impressions than less diagnostic traits of equivalent likeability. Moreover, we were able to isolate three underlying factors of trait diagnosticity. In addition to negative (vs. positive) valence and extremity, a trait's high average distance to the other traits within the sample (in accordance with the density-model framework) provided a third source of diagnosticity. The research design of all experiments involved pre-selected population sets from which samples were drawn randomly. Since each participant received samples from all sets and since the different population sets were constructed to carry orthogonal combinations of diagnosticity determinants, the research design allowed us to precisely predict the resulting group impressions from orthogonal diagnosticity predictors within participants.

While the entire pattern corroborates the usefulness and fertility of a sampling-theoretical approach to impression formation (Norton et al., 2007; Prager et al., 2018; Ullrich et al., 2013), the main original contribution of our research consisted in the delineation of self-truncation effects, thereby offering a completely novel perspective on inter-group judgment. Whereas the law of large numbers and Bayesian updating of flat priors predicts that the same proportion (e.g., proportion of positive traits, correct student responses, or favorable consumer ratings) provides stronger evidence when observed in a larger than in a smaller sample (Bernoulli, 1713; Tversky & Kahneman, 1971), self-truncated sampling removes this large-sample advantage. Thus, 12 positive votes in a sample of $n = 16$ provides stronger evidence for a positive outcome rate of 75% than the same proportion of 3 positive outcomes in $n = 4$, as long as sample size is treated as independent variable. However, self-truncation causes a notable reversal of the positive relation between evidence strength and size of a sample. When the first few observations happen to reflect a strong and regular trend (e.g., 3 or 4 positive outcomes in $n = 4$), early truncation produces polarized and homogeneous samples that can be expected to trigger strong and conflict-free judgments. In the absence of such a stochastic primacy effect, samples of increasing size are quite unlikely to reach similarly extreme proportions and reduced variance as is possible in early phases of growing samples. Thus, because self-truncated samples can be expected to remain small if they exhibit strong and conflict-free patterns but become large when the initial evidence is weak and conflict-prone, it is no wonder that self-truncated samples tend to convey stronger evidence when they are small rather than large.

Note that this strong reversal from a positive correlation between sample size and evidence strength with experimenter-determined n to a negative correlation with self-truncated n is possible because n is no longer an independent variable but dependent on the judge's primacy impression of strong evidence that justifies early truncation. Conversely, a large self-truncated n is reflective of judge's appraisal that a samples started with weak evidence that did not justify earlier truncation.

Two important implications of self-truncation effects deserve to be emphasized. First, although the evidence for a strong reversal from a positive to a negative impact of sample size on evidence strength has been replicated in various experiments and substantiated in simulation studies (Fiedler et al., 1999; Fiedler et al., 2010; Prager & Fiedler, 2021), its relevance for political, economic, or health-related judgments and decisions is hardly ever recognized. For instance, protocols of democratic decision groups do not reveal whether a discussion underlying a consequential decision was self-truncated or externally determined, or whether consumer choices were informed by self-truncated or externally truncated sampling. In all these domains, we continue to presuppose that more extensive information acquisition and more careful advice taking produces more accurate decisions, if only to justify information costs.

Secondly, it is obvious that sample truncation decisions (i.e., the decisions to stop a sequentially unfolding sample) are similarly sensitive to stimulus diagnosticity as the final judgments. The evidence from the present investigation provides strong support for this notion. The enhanced diagnosticity of negative, extreme, and low within-sample-distance traits (compared to less diagnostic positive, moderate, and high-distance traits) not only led to stronger and more homogenous final impression judgments (manifested in stronger impression strength J and homogeneity H scores), but also enabled earlier truncation ($\sqrt{n}$) leading to exaggerated sample estimates. As a consequence, the dominant impact of trait diagnosticity on truncation served to amplify the diagnosticity effect on the resulting impressions, as manifested in regularly negative correlations between sample size and impression strength $r_{\sqrt{n},J}$ and between sample size and homogeneity $r_{\sqrt{n},J}$. The process model depicted in Figure 3, which guided the present research, explicates the co-existence of a direct influence of sampled stimulus properties on group impressions and an indirect effect mediated by truncation effects. Empirical support for this two-fold influence can be found in systematic (negative) relations between diagnostic trait properties (negativity, extremity, and distance) on one hand and judgment strength (J), homogeneity (H) and negative $r_{\sqrt{n},J}$ and $r_{\sqrt{n},H}$ correlations on the other hand.

One appealing implication of this analysis is that our sampling approach offers a sufficient and parsimonious account of both major phenomena of inter-group research. Granting that experienced samples are typically smaller for out-groups than for in-groups, the reported findings afford a sufficient account of out-group homogeneity (H) and out-group polarization (J). Because negative behavior (in the communion domain; Reeder & Brewer, 1979) is more diagnostic than positive behavior, enhanced homogeneity and polarization also implies out-group derogation. Consistent with this sample size account of inter-group judgments, pertinent research has shown that out-group polarization and homogeneity are ameliorated or even reversed from minority perspectives (when out-group sample size is relatively high; Simon Brown, 1987) or when asymmetric social contact serves to reduce the sample-size difference between in-groups and out-groups (Wagner et al., 2006).

For the sake of theoretical clarity, it seems appropriate to note that we do not state that inter-group judgments can be reduced to variation in sample size reflecting variation in diagnosticity. In a multi-causal world, inter-group relations and inter-group judgments are presumably sensitive to many different causal impacts, including real conflicts, resentments, cultural and linguistic influences, closeness of social inter-connections, and the distribution of resources. Rather than propagating a necessary condition supposed to underlie all inter-group biases, we argue, and have provided cogent evidence to demonstrate, that unequal samples size constitutes a sufficient condition for the most prominent biases reported in inter-group literature. We believe that this demonstration is very useful for progress in future research, which should try to disentangle inter-group effects that go beyond the basic sampling effects that were the focus of the present article.

References

Bernoulli, J. (1713). *Ars conjectandi: Opus posthumum.* Thurnisii.

Brewer, M. B. (1993). Social identity, distinctiveness, and in-group homogeneity. *Social Cognition*, *11*(1), 150–164.

Brockbank, E., Holdaway, C., Acosta-Kane, D., & Vul, E. (in press). Sampling data, beliefs, and actions. In K. Fiedler, P. Juslin, & J. Denrell (Eds.), *Sampling in judgment and decision making.* Cambridge University Press.

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science*, *3*, 14–25.

De Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, *7*(1), 1–68.

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review*, *112*(4), 951–978.

Denrell, J., & Le Mens, G. (2007). Interdependent sampling and social influence. *Psychological Review*, *114*(2), 398–422.

Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science*, *12*(5), 523–538.

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology*, *87*(3), 293–311.

Fiedler, K., Kemmelmeier, M., & Freytag, P. (1999). Explaining asymmetric inter-group judgments through differential aggregation: Computer simulations and some new evidence. *European Review of Social Psychology*, *10*(1), 1–40.

Fiedler, K. (2000). Beware of samples! a cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*(4), 659–676.

Fiedler, K., Renn, S.-Y., & Kareev, Y. (2010). Mood and judgments based on sequential sampling. *Journal of Behavioral Decision Making*, *23*(5), 483–495.

Fiedler, K., & Wänke, M. (2009). The cognitive-ecological approach to rationality in social psychology. *Social Cognition*, *27*(5), 699–732.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

Gidron, D., Koehler, D. J., & Tversky, A. (1993). Implicit quantification of personality traits. *Personality and Social Psychology Bulletin*, *19*(5), 594–604.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments Computers*(4), 381–386.

Haldane, J. (1945). A labour-saving method of sampling. *Nature*, *155*, 49–50.

Kareev, Y., Arnon, S., & Horwitz-Zeliger, R. (2002). On the misperception of variability. *Journal of Experimental Psychology: General*, *131*(2), 287–297.

Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology*, *110*(5), 675–709.

Konovalova, E., & Le Mens, G. (2020). An information sampling explanation for the in-group heterogeneity effect. *Psychological Review*, *127*(1), 47–73.

Le Mens, G., & Denrell, J. (2011). Rational learning and information sampling: On the 'naivety' assumption in sampling explanations of judgment biases. *Psychological Review*, *118*(2), 379–392.

Leibniz Institut Für Deutsche Sprache. (2014). Datenbank für gesprochenes Deutsch [spoken german database].

Leibniz Institut Für Deutsche Sprache. (2017). Deutsches Referenzkorpus [german reference corpus].

Linville, P. W., & Fischer, G. W. (1993). Exemplar and abstraction models of perceived group variability and stereotypicality. *Social Cognition*, *11*(1), 92–125.

Linville, P. W., & Jones, E. E. (1980). Polarized appraisals of out-group members. *Journal of Personality and Social Psychology*, *38*(5), 689–703.

Linville, P. W., Fischer, G. W., & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: Empirical evidence and a computer simulation. *Journal of Personality and Social Psychology*, *57*(2), 165–188.

Norton, M. I., Frost, J. H., & Ariely, D. (2007). Less is more: The lure of ambiguity, or why familiarity breeds contempt. *Journal of Personality and Social Psychology*, *92*(1), 97–105.

Park, B., & Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. *Journal of Personality and Social Psychology*, *42*(6), 1051–1068.

Prager, J., & Fiedler, K. (2021). Forming impressions from self-truncated samples of traits - interplay of Thurstonian and Brunswikian sampling effects. *Journal of Personality and Social Psychology*, *121*(3), 474–497.

Prager, J., Krueger, J. I., & Fiedler, K. (2018). Towards a deeper understanding of impression formation-new insights gained from a cognitive-ecological perspective. *Journal of Personality and Social Psychology*, *115*(3), 379–397.

Quattrone, G. A., & Jones, E. E. (1980). The perception of variability within in-groups and out-groups: Implications for the law of small numbers. *Journal of Personality and Social Psychology*, *38*(1), 141–152.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, *86*(1), 61–79.

Rothbart, M., & Park, B. (1986). On the confirmability and disconfirmability of trait concepts. *Journal of Personality and Social Psychology*, *50*(1), 131–142.

Sharpe, W. F., Goldstein, D. G., & Blythe, P. W. (2000). The distribution builder: A tool for inferring investor preferences.

Simon, B., & Brown, R. (1987). Perceived intragroup homogeneity in minority-majority contexts. *Journal of Personality and Social Psychology*, *53*(4), 703–711.

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology*, *52*(4), 689–699.

Thorndike, E. L. (1927). The law of effect. *The American Journal of Psychology*, *39*, 212–222.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110.

Ullrich, J., Krueger, J. I., Brod, A., & Groschupf, F. (2013). More is not less: Greater information quantity does not diminish liking. *Journal of Personality and Social Psychology*, *105*(6), 909–920.

Unkelbach, C., Fiedler, K., Bayer, M., Stegmüller, M., & Danner, D. (2008). Why positive information is processed faster: The density hypothesis. *Journal of Personality and Social Psychology*, *95*(1), 36–49.

Wagner, U., Christ, O., Pettigrew, T. F., Stellmacher, J., & Wolf, C. (2006). Prejudice and minority proportion: Contact instead of threat effects. *Social Psychology Quarterly*, *69*(4), 380–390.

Yzerbyt, V., Castano, E., Leyens, J. P., & Paladino, M. P. (2000). The primacy of the ingroup: The interplay of entitativity and identification. *European Review of Social Psychology*, *11*(1), 257–295.

Appendix A. Pre-Study: Ratings on Natural Social Groups

Table A1. Naturally occurring social groups adapted from Koch et al. (2016) with average ratings on knowledge (2 items averaged), likeability and within-group homogeneity.

group label (German original)	group label (English translation)	knowledge	likeability	homogeneity
Ärzte	physicians	0.50	0.62	0.42
Arbeiter	workers	0.57	0.61	0.40
Arbeitslose	unemployed	0.27	0.45	0.33
Auszubildende	trainees	0.41	0.59	0.35
Autofahrer	car drivers	0.79	0.44	0.24
Buddhisten	Buddhists	0.20	0.62	0.51
Christen	Christians	0.77	0.53	0.36
Fahrradfahrer	bicycle drivers	0.79	0.63	0.38
Fußballspieler	soccer players	0.36	0.42	0.55
Hipster	hipsters	0.39	0.44	0.64
Homosexuelle	homosexuals	0.53	0.70	0.31
Konservative	conservatives	0.50	0.33	0.63
Künstler	artists	0.39	0.66	0.37
Lehrer	teachers	0.69	0.60	0.32
Manager	managers	0.29	0.41	0.54
Musiker	musicians	0.49	0.70	0.36
Muslime	Muslims	0.39	0.52	0.34
Obdachlose	homeless	0.17	0.47	0.42
Politiker	politicians	0.40	0.43	0.44
Punks	punks	0.18	0.43	0.61
Reiche	rich	0.45	0.42	0.42
Rentner	retiree	0.58	0.53	0.37
Schüler	pupils	0.75	0.55	0.28
Selbstständige	freelancers	0.45	0.59	0.34
Singles	singles	0.77	0.62	0.17
Städter	town people	0.72	0.60	0.38
Studenten	(university) students	0.90	0.74	0.27
Veganer	vegans	0.57	0.57	0.56

Appendix B. Experiment 4: Contrast Analysis

Table B1. Contrasts used for the analyses on valence and frequency in Experiment 4. The four columns relate to the four population sets applied in the experiment.

	Negative natural	Positive natural	Negative reversed	Positive reversed
Contrast 1: valence	$-\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$
Contrast 2: frequency	$\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$
Contrast 3: natural vs. reversed	$\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$

Note. Natural sets indicate that the naturally occurring correlation of higher frequency for positive than negative is preserved, which is reversed for the respective two other sets.

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

**FAKULTÄT FÜR VERHALTENS-
UND EMPIRISCHE KULTURWISSENSCHAFTEN**

Promotionsausschuss der Fakultät für Verhaltens- und Empirische Kulturwissenschaften der Ruprecht-Karls-Universität Heidelberg / Doctoral Committee of the Faculty of Behavioural and Cultural Studies of Heidelberg University

Erklärung gemäß § 8 (1) c) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften / Declaration in accordance to § 8 (1) c) of the doctoral degree regulation of Heidelberg University, Faculty of Behavioural and Cultural Studies

Ich erkläre, dass ich die vorgelegte Dissertation selbstständig angefertigt, nur die angegebenen Hilfsmittel benutzt und die Zitate gekennzeichnet habe. / I declare that I have made the submitted dissertation independently, using only the specified tools and have correctly marked all quotations.

Erklärung gemäß § 8 (1) d) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften / Declaration in accordance to § 8 (1) d) of the doctoral degree regulation of Heidelberg University, Faculty of Behavioural and Cultural Studies

Ich erkläre, dass ich die vorgelegte Dissertation in dieser oder einer anderen Form nicht anderweitig als Prüfungsarbeit verwendet oder einer anderen Fakultät als Dissertation vorgelegt habe. / I declare that I did not use the submitted dissertation in this or any other form as an examination paper until now and that I did not submit it in another faculty.

Vorname Nachname / First name Family name	Johannes Prager
Datum / Date	02.12.2021
Unterschrift / Signature	J. P