Developing an assessment of epistemic trust: a research protocol

Paul Schröder-Pfeifer, Alessandro Talia, Jana Volkert, Svenja Taubner

Institute of Psychosocial Prevention, Heidelberg University Hospital, Heidelberg, Germany

ABSTRACT

Epistemic trust (ET) describes the willingness to accept new information from another person as trustworthy, generalizable, and relevant. It has been recently proposed that a pervasive failure to establish epistemic trust may underpin personality disorders. Although the introduction of the concept of ET has been inspiring to clinicians and is already impacting the field, the idea that there may be individual differences in ET has yet to be operationalized and tested empirically. This report illustrates the development of an Epistemic trust assessment and describes the protocol for its validation. The sample will include 60 university students. The Trier Social Stress Test for Groups will be administered to induce a state of uncertainty and stress, thereby increasing the relevance of information for the participants. The experiment will entail asking information from the participants about their performance and internal states during a simulated employment interview, and then tracking how participants are able to revise their own judgments about themselves in light of the feedback coming from an expert committee. To control for social desirability and personality disorder traits, the short scale for social desirability (Kurzskala Soziale Erwünschtheit-Gamma) and the Inventory of Personality Organization are utilized. After the procedure, the participants will complete an app-based Epistemic trust questionnaire (ETQ) app. Confirmatory Factor Analysis will be utilized to investigate the structure and dimensionality of the ETQ, and ANOVAs will be used to investigate mean differences within and between persons for ET scores by item category. This study operationalizes a newly developed ET paradigm and provides a framework for the investigation of the theoretical assumptions about the connection of ET and personality functioning.

Key words: Epistemic trust; Experiment; Operationalization.

Correspondence: Paul Schröder-Pfeifer, Institute of Psychosocial Prevention, Heidelberg University Hospital, Bergheimer Str. 54, 69115 Heidelberg, Germany.
Tel.: +49.6221.56.38504 - Fax: +49.6221.56.4702.
E-mail: Paul.Schroeder-Pfeifer@med.uni-heidelberg.de

Citation: Schröder-Pfeifer, P., Talia, A., Volkert, J., & Taubner, S. Developing an assessment of epistemic trust: a research protocol. *Research in Psychotherapy: Psychopathology, Process and Outcome, 21*(3), 123-131. doi: 10.4081/ripppo.2018.330

Contributions: all the authors participated in designing the experiment. PS-F did the literature review and first draft of the paper; AT and PS-F provided the introduction; JV and ST revised and edited the manuscript.

Conflict of interest: the authors declare no potential conflict of interest.

Funding: the German psychoanalytic society (Deutsche Psychoanalytische Gesellschaft) provided funding to reimburse the study participants.

Conference presentation: the first draft of the experimental design was presented at the Society for Psychotherapy Research Meeting, Toronto 2017; International Society for the Study of Personality Disorders Meeting, Heidelberg 2017; Deutsche Psychoanalytische Gesellschaft Yearly Meeting 2018; and Society for Psychotherapy Research Meeting, Amsterdam 2018.

Received for publication: 22 August 2018.
Revision received: 20 November 2018.
Accepted for publication: 26 November 2018.

Research in Psychotherapy:
Psychopathology, Process and Outcome 2018; 21:123-131
doi:10.4081/ripppo.2018.330

Introduction

In the past few years, an important shift has occurred in Fonagy, Luyten, Allison, and Campbell views on psychopathology (Fonagy, Luyten, Allison, & Campbell, 2017). Previously (2004), these authors argued that the capacity to reflect on mental states underlying behavior (*i.e.* the capacity to mentalize) is a developmental achievement that arises out of secure attachments, and that mentalizing and secure attachment constitute a source of resilience against psychopathology (Fonagy, 2004). More recently, however, Fonagy, Luyten, and Allison (2015) have proposed that it is disruptions in early social communication - rather than in early attachments or mentalizing *per se* - that lead to subsequent vulnerabilities for psychopathology. Drawing among others from Csibra and Gergely's Natural Pedagogy (Csibra & Gergely, 2009), and from Sperber and Wilson's Relevance Theory (Sperber et al., 2010; Wilson & Sperber, 2012), Fonagy et al. have built the case that psychopathology, insecure attachment, and impaired mentalizing are all linked because they are associated with difficulties in trusting the relevance and generalizability of intentional communication (Fonagy et al., 2017). They refer to this capacity with the term "epistemic trust", and they view its recovery as lying at the heart of any effective psychotherapy.

While these novel views have started to impact clinical and theoretical work (Bateman & Fonagy, 2016; Holmes & Slade, 2017), there is still very little empirical work to

support them. In fact, while the concept of ET has inspired a growing empirical literature in developmental psychology (*e.g.* Corriveau & Harris, 2009; Hagá & Olson, 2017; Harris & Corriveau, 2011), the study of the concept in adolescents, adults and (in particular) clinical populations is still in its infancy. In particular, there is no valid measure of ET available today for adolescents or adults.

The current study attempts to fill this research gap by devising an assessment of epistemic trust (ET) that attempts to translate the theoretical assumptions of the clinically informed ET literature into a valid experimental paradigm. Most of the work in this field up today is theoretical in nature, and further developments in this area of research are likely to depend on methodological advancements related to the measurement of ET. After a brief review of the theoretical framework and empirical literature for this study, in the following we describe the development of our assessment of ET and a protocol for its validation.

Epistemic trust and epistemic vigilance

Learning involves, by definition, some kind of generalization of the import of new information that is learnt on a specific occasion (*i.e.* at a specific time and in a specific place) to novel instances where the information can be used for a different goal or in a different context. Theories of learning usually argue that such generalization relies on statistical procedures that sample multiple episodes (Csibra & Gergely, 2009). Humans, however, can acquire generic knowledge from a single instance in which they gain new information, *i.e.* through intentional communication with a trusted person. For example, from many repeated observations, one may learn that a particular series of movements leads to having one's shoes laced. Yet if the person (*e.g.*, a parent) who is performing those movements does not merely perform the sequence of actions, but performs it *manifestly* for their addressee (*e.g.*, a child) by clearly indicating that this is a demonstration presented to them specifically, they will learn significantly more from the same action than they would from simply observing how it is performed. In other words, by providing information *ostensively* (*i.e.* by indicating an intent to communicate, Sperber & Wilson, 1995), it may suffice one or two demonstrations from a trusted other (*i.e.* a parent) about *e.g.*, "how one ties shoe laces" to transmit information reliably.

Mammal species have developed mechanisms to protect themselves from deception; similarly, humans depend to a large extent on communication with others, which leaves them open to the risk of being misinformed, sometimes intentionally. To ensure that communication remains advantageous, humans must possess a suite of mechanisms for epistemic vigilance (Sperber et al., 2010). However, the human capacity to acquire from others information that has social and cultural significance may rely on a special kind of trust that may be characteristic of the human species.

Csibra & Gergely (2009) have made the claim that human communication is adapted to allow the transmission of generic knowledge between individuals in at least two distinct ways. First, human infants are sensitive by default to ostensive signals that indicate that they are being addressed. Ostensive cues like eye contact, motherese and marked mirroring prepare the interlocutor for information specifically *relevant* to them, thereby increasing the chance of the information being accepted and generalized to other circumstances, interaction partners and situations (Csibra & Gergely, 2009; Egyed, Kiraly, & Gergely, 2013). Second, humans may be biased to interpret ostensive communication as conveying information that is generalizable – *i.e.* have ET.

Epistemic trust, psychopathology, and psychotherapy

Fonagy et al. have drawn from these views to argue about the importance of ET in psychopathology and psychotherapy. ET within an individual is thought to develop in early attachment relationships with primary caregivers (Csibra & Gergely, 2009; Fonagy et al., 2015). In this perspective, personality disorder is seen as descending from a failure to establish ET in early relationships, and identifiable by persistent problems in communication that reveal a lack of trust in interpersonally transmitted information (Allison & Fonagy, 2016; Fonagy et al., 2015; Fonagy & Allison, 2014).

A healthy ET can be described as the capacity to exert appropriate vigilance in the face of possible deceit while maintaining general trust in interpersonally transmitted information (Sperber et al., 2010). On the other hand, the capacity for ET of an individual can be limited in one of two ways. First, an individual might be epistemically hypervigilant (Sperber et al., 2010) or petrified (Fonagy & Allison, 2014), unable to accept information from the outside world, and rigid in their mental states and in behavior. Second, an individual might be epistemically naïve (Sperber et al., 2010), which might lead to a predisposition to being more easily deceived and naïve behavior.

For example, patients with a borderline personality disorder (BPD) have been found to systematically over-attribute hostile intentions to other people (Nicol, Pope, Sprengelmeyer, Young, & Hall, 2013), over-interpret motives of other people (Sharp et al., 2011; Sharp et al., 2013), and broadly speaking misattributing mental states (*e.g.* Daros, Uliaszek, & Ruocco, 2014; Matzke, Herpertz, Berger, Fleischer, & Domes, 2014). Research suggests that patients with BPD consistently perceive the reason for someone's behavior as threatening or at least malevolent and therefore disregard information provided by their social interaction partners, consistent with their view of the social world being generally malevolent. This phenomenon is not only found in BPD but also in other personality disorders (*e.g.* Bateman & Fonagy, 2016; Beck, Davis, & Freeman, 2016; Schnell & Herpertz, 2018). It translates into a rigidity that hinders the normally ongoing process of updating the

self (beliefs about the world and oneself) based on information from the social environment.

ET has also been discussed as a general mechanism of change in psychotherapy. In psychotherapy, interpersonal processes like empathy, mentalization, and the therapeutic alliance may be considered to function as ostensive cues (Csibra & Gergely, 2009; Fonagy & Allison, 2014). The importance assigned to ET seems compatible with most theories of psychotherapy (*e.g.*, cognitive, psychoanalytic, humanistic) because it tackles a human learning process addressed in any therapeutic intervention: the capacity to learn from experience. The feeling of being understood, of finding oneself accurately represented in the mind of another, rekindles ET and thus might reestablish trust in social learning. This is of central importance for the therapy of individuals with epistemic petrification, which normally experience a sense of isolation from the social world due to communicative pathways with others being essentially severed (Fonagy et al., 2015). Over time, in a benevolent social environment, this may also generalize beyond the therapeutic setting as it enables increasingly accurate interpretation of other's mental states (Fonagy et al., 2015; Fonagy & Allison, 2014).

Previous research

While conceptual work on ET promises to advance our understanding of developmental psychopathology and psychotherapy, there is a need for a valid instrument that assesses ET in adolescents and adults and therewith provides an empirical validation for this clinical theory. In devising our ET instrument we have drawn from previous experimental work carried on young children (Corriveau & Harris, 2009; Egyed et al., 2013). In the following paragraphs we describe these earlier studies and then present how we developed our instrument to study ET in adults. Egyed in his experiment (Egyed et al., 2013) sets out to study the mechanism of ET in toddlers. In Egyed's experiment n=48 toddlers aged 18 months were seated across a table with an experimenter. On the table in between the toddler and the experimenter were placed two objects, one blue object to the right and one orange object to the left. In the first condition, the experimenter first smiled at the blue object and then looked disgusted towards the orange object. The experimenter then left the room and a second person entered and asked the toddler to hand her one of the objects. In 31% of the cases, the toddler handed the object preferred by the experimenter. In contrast, in the second condition, where the experimenter established ostensive contact with the toddler by smiling and eyecontact, the toddler handed the second person the object preferred by the experimenter in 69% of the cases. It can be assumed that the toddler generalized the information regarding the preference beyond the dyadic interaction.

The experiment by Corriveau and Harris (2009) with 147 young children at the age of four to five years works similarly. The children were presented with pictures of fantasy animals and had to choose one of two labels for the animals, one provided by the child's mother, the other by a stranger. The fantasy animals were either completely unfamiliar or hybrid animals that were made up of two different animals in proportions of 50/50 or 75/25. With the unfamiliar animals and the 50/50 ones, the mother and the stranger supplied different, yet fitting labels. For the 75/25 animals, the mother labeled the part of the animal that corresponded to the 25% part while the stranger supplied the label that corresponded to the 75% part. In this experiment, epistemic vigilance would correspond to the children choosing the label supplied from the mother for the unfamiliar and 50/50 conditions, and the label of the stranger for the 75/25 condition.

Both experiments assess ET by measuring how new information is processed by the child. For the information to actually be processed by ET, the information has to be relevant Sperber et al. (2010). Gilbert et al. were able to show that information that has no specific relevance to the subject is automatically accepted as truthful, but is not internalized (Gilbert, Krull, & Malone, 1990; Gilbert, Tafarodi, & Malone, 1993). Non-salient information is not relevant for the self on a conscious or unconscious level, accordingly, there is no risk associated in accepting it, as the information is not considered relevant at any point in the future. At the same time, while keeping the processing cost at a bare minimum, it might be evolutionary optimal to accept non-relevant information as true if it was not merely uttered but asserted, as assuming the information was false would require the individual to question the legitimacy of the assertion.

While it is relatively easy to experimentally establish relevance with young children, it is more difficult to create salient material for adults, who have already formed interests and knowledge. For new information to achieve relevance in the context of existing beliefs, one of three conditions has to be met (Sperber & Wilson, 1995): i) Implications arise taking the new information and contextual beliefs together as premises, which are not derivable from neither the context nor the new information alone. These implications are then accepted as new beliefs. ii) The individual has to adjust their confidence in contextually activated beliefs when taking in the new information. iii) The individual's prior beliefs might contradict the new information. Either the new information has to be rejected or the existing beliefs have to be remodeled accordingly (Sperber & Wilson, 1995).

A further challenge is that the majority of experiments that aim to assess ET with children restrict themselves to presenting to participants declarative information (*e.g.* Corriveau & Harris, 2009; Egyed et al., 2013; Hagá & Olson, 2017). While declarative information has the advantage of establishing the *correct* answer to statements and questions, it may fail to touch on the more socially focused aspects of ET in which *correctness* of inherently subjective information like feedback on a performance has to be established within social interactions.

In sum, the relevance of ET in the field of psychotherapy research has substantially grown in recent years. Yet, to the best of our knowledge, there is no valid measure of ET available for adolescents or adults although some are in development (*e.g.* Luyten, 2017; Nolte, 2017). Accordingly, this study aims to develop an experimental paradigm for the assessment of ET that closely relating to its theoretical basis.

Materials and Methods

Participants

Participants will be students of the University of Heidelberg who have voluntarily signed up to participate in studies via an online study participation platform. Students are notified about the platform by e-mail when they first sign up to university. Inclusion criteria are age above 18, able to provide informed consent, and are fluent in the German language. 2424 registered students at the time of the sighting of the recruitment pool were filtered according to the inclusion criteria and the recruited sample was selected randomly by a computer tool build into the platform from a pool of 1737 eligible students (Figure 1).

Development of the epistemic trust assessment

Building on the ET experiments designed by Egyed et al. (2013) or Corriveau et al. (2009) with young children, we designed the epistemic trust assessment (ETA) to control and observe the content and amount of information passed to an individual and the degree to which the individual internalizes and generalizes that information, this way providing an indirect estimate of ET. Based on the results from Gilbert et al. (Gilbert et al., 1990; Gilbert et al., 1993), and also previous tries at operationalizing ET (Luyten, 2017; Nolte, 2017), the ETA is developed with a focus on the relevance of information passed to the participants. Furthermore, as research from business, organizational and cognitive psychology suggests that individuals experiencing stress are more prone to gathering information from external sources to combat the uncertainty resulting from the stress (Driskell & Salas, 1991; Eysenck, Derakshan, Santos, & Calvo, 2007; Starcke & Brand, 2012), the ETA was devised for use in combination with an artificial stressor, the Trier Social Stress Test for Groups (TSST-G) (Dawans, Kirschbaum, & Heinrichs, 2011), to increase the relevance of the information. According to the theoretical conceptualization of ET outlined above, establishing salience of the information for the participants is of outmost importance, as irrelevant information has no consequences for the individual and activation of ET is not necessary.

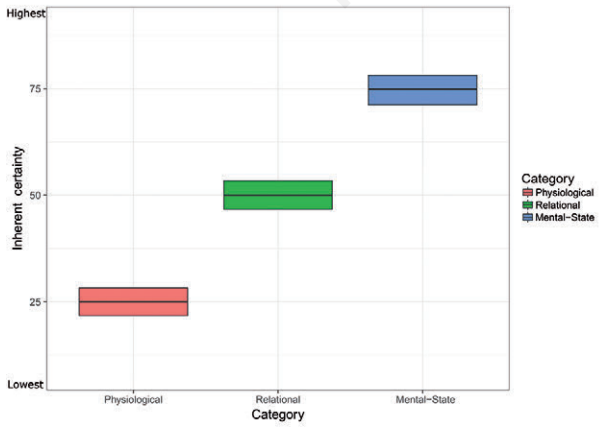

Figure 1. Categories of epistemic trust statements: physiological, relational, mental-state and their inherent certainty (low, medium, high).

In sum, this study aimed to design an experiment utilizing the TSST-G to provide both relevant information to communicate to the participants as well as a context and increased relevance by virtue of providing a stressor. We set out to answer the question, whether or not an experiment can be devised that measures ET and deviations from ET by assessing if participants generalize information supplied to them, given different levels of inherent certainty nested in specific statements. We hypothesize that information can be classified in categories of relative certainty. For the development of the ETA, we differentiated three categories of information that are distinct in terms of their degree of certainty: i) information regarding one's own physiological state (low inherent certainty), ii) regarding relational states (medium inherent certainty), and iii) regarding one's mental state (high inherent certainty). These categories describe three different levels of certainty during the encounter between participant and TSST-G *expert* committee. We assume that specific information about one's own physiological state should be opaque to the individual, and thus have a low inherent certainty. As such, a feedback statement from the expert committee on the individual's heart rate "At the moment your heart rate is around 90 beats per minute." should be difficult to evaluate without the use of technological aides, making questions on physiological states prone to be influenced by feedback. With regard to information on one's mental states is characterized by a high inherent certainty. Assuming that the individual has privileged access to one's mental states, this information should be characterized as high inherent certainty and not be influenced by information from external feedback. Information about relational states can be considered to be of medium inherent certainty as all partners in an interpersonal encounter are considered to have both individual and shared intrapersonal and interpersonal subjective information about the relationship. An individual may have his own judgment on how he is perceived from the outside, but cannot be certain. Consequently, statements regarding relational states should be influenced in a medium way by feedback (Figure 1, Table 1).

Hypotheses

Primary hypothesis

The main hypothesis is that participants adjust their certainty post-feedback according to statement categories and not independent of them. This is assuming a normative sample of participants with healthy epistemic vigilance.

$H0_1$: The participants adjust their certainty post-feedback independent of statement category.
$H1_1$: The participants adjust their certainty post-feedback dependent on category, with most change in the physiological category and least change in the mental states category.

Secondary hypothesis

The secondary hypothesis addresses the relationship between BPD traits and ET. Fonagy et al. (Fonagy et al., 2015; Fonagy & Allison, 2014) conceptualize BPD with the loss of epistemic vigilance tending towards epistemic hypervigilance or equivalent *epistemic petrification*. Accordingly, it is hypothesized that participants with BPD traits adjust their judgments post-feedback significantly less than participants without BPD traits.

$H0_2$: Participants with BPD traits according to the Inventory of Personality Organization (IPO-16) cut-off values adjust their certainty post-feedback the same as participants without BPD traits.
$H1_2$: Participants with BPD traits adjust their certainty post-feedback by significantly less then participants without BPD traits.

Assessment of epistemic trust

Epistemic trust questionnaire

The epistemic trust questionnaire (ETQ) is a self-report questionnaire in app form for the indirect assessment of ET following the ETA. The questionnaire consists of three parts. In the first part, the participants have to rate, according to the 3 certainty categories, their physiological state, their mental states during the TSST-G, and their relational state (*e.g.*, i) "Do you think, your blood pressure (in mmHg) was high or low during the experiment?", ii) "Were you bored during the interview?", iii)"Do you think you came across as motivated?"), and, more importantly, how certain they are in making their judgement. In the second part, the participants are presented with a standardized, computer-generated feedback they think was given to them by the committee, on all of the statements they answered during step one. Finally, in the third step, the participants are asked to re-rate their certainty for the items answered during the first step, taking into account the new information. The items in the first and third step all entail a rating of certainty on a scale of 0 to 100 as well as a binary rating of valence ("Yes/No", "High/Low", *etc*; Figure 2).

The feedback is computer-generated in order to be standardized and is in accordance with the participant's valence rating in exactly half of the questions, as not to introduce a bias on over- or under-agreement. The ET score is operationalized as the difference in certainty from step one to step three, relative to item category. Epistemic vigilance is associated with big changes towards more certainty in the physiological items, medium changes in either direction in the relational items, and no change or small changes in either direction in the mental states items. This operationalization exemplifies epistemic vigilance as a construct of balance that should prompt individuals to internalize and accept information where it is meaningful for them and certainty about their judgment should be low (low certainty item category physiological state). Accordingly individuals should distrust and therefore not internalize information where it is unlikely to meaningfully update their prior knowledge (high certainty item category mental state). Epistemic hypervigiliance is

Table 1. Inherent certainty categories, example items and predisposition to change of the epistemic trust questionnaire.

Category	Example item	Inherent certainty	Predisposition to change
Physiological	"Was your pulse, on average, below or above 97 during the experiment?"	Low	High
Relational	"Do you think you came across as friendly or unfriendly during the experiment?"	Medium	Medium
Mental-State	"Did you feel anxious during the experiment?"	High	Low

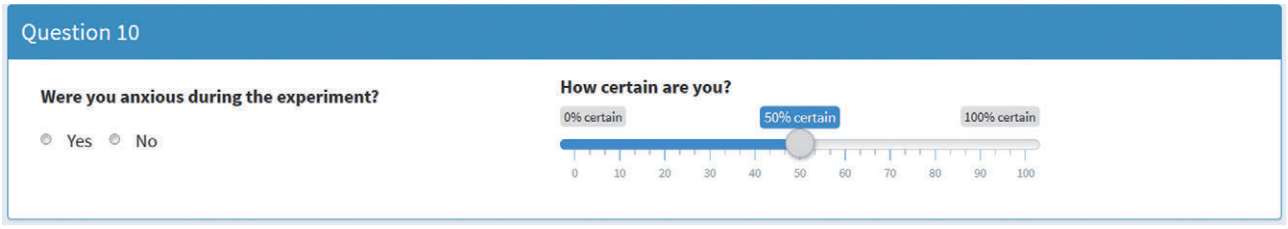

Figure 2. Sample question from the epistemic trust questionnaire.

associated with no or small changes in either direction independent of item category, while epistemic naïveté is associated with big changes towards more certainty independent of item category.

A possible effect known from research on metacognitive phenomenon that might interfere with our hypothesis on how ET is operationalized by the experiment is the so called hypercorrection effect (*e.g*. Butterfield & Metcalfe, 2001; Metcalfe & Finn, 2012). This effect describes a tendency to more easily correct apparently wrong statements that were of high prior certainty as opposed to low prior certainty. This might lead to participants overcorrecting statements with high inherent certainty, such as from the relational and mental states category. However, while this effect has not yet been thoroughly examined for non-declarative information, and research suggests that participants have to be relatively sure that the alternative statement provided to them is correct feedback (Metcalfe & Finn, 2011). In the face of non-declarative information like the feedback provided by the committee in this study, it seems unlikely that this effect applies for any of the categories except for the physiological information, since both relational and mental state information is inherently subjective and can thus never be entirely *correct*.

Social stress test

The Trier Social Stress Test for Groups (TSST-G) (Dawans, Kirschbaum, & Heinrichs, 2011) is a standardized experiment for the reliable induction of moderate social stress (Dawans et al., 2011). The TSST-G is the group version for up to six participants of the original paradigm by Foley and Kirschbaum (2010). The six participants take part in a fabricated job interview combined with an arithmetic task in front of a panel of *experts*. During the interview and the arithmetic tasks, participants cannot see each other, are instructed that they can be called upon at any time in a random order and are being filmed by two cameras. The *expert* panel is instructed to stress the participants by interrupting participants during the interview with questions, if they speak too fluent or too slow as well as prompting them to calculate faster. One *expert* member is the active one, interrupting the participants and asking questions, while the other is appearing to take notes on a laptop for the appearance that data actually utilized. This is a slight modification of the original procedure where the other *expert* member is completely passive. The TSST-G has been shown to reliably induce a robust increase in the activation of the hypothalamic-pituitary-adrenal stress system (Boesch et al., 2014; Kirschbaum, Kudielka, Gaab, Schommer, & Hellhammer, 1999; Kirschbaum, Pirke, & Hellhammer, 1993; Leder, Hausser, & Mojzisch, 2013).

Assessment of social desirability

The Short Scale Social Desirability-Gamma (Kurzskala Soziale Erwünschtheit-Gamma; KSE-G) (Kemper, Beierlein, Bensch, Kovaleva, & Rammstedt, 2012) is an economic measure for the assessment of social desirable behavior (Paulhus, 2015). The scale measures aspects of social desirability associated with a moralistic bias to deny unwanted impulses and to appeal unrealistically positive in the eyes of others. The participants rate six items describing social behavior (*i.e*. "When in an argument, I always stay factual and objective") on a 5-point Likert scale ranging from "does not apply at all" to "applies fully". The authors report satisfactory internal consistency and high factorial and content validity of the instrument (Kemper et al., 2012).

Assessment of personality functioning

The 16-Item-Version IPO-16 (Zimmermann et al., 2013) is a self-report measure to assess personality functioning based on Kernberg's model of borderline personality organisation with regard to identity diffusion, primitive psychological defenses and reality testing. The items are rated on a 5-point Likert scale ranging from "never applies" to "always applies". The authors report good internal constancy (α=.85) and good discriminant, as well as convergent validity (Zimmermann et al., 2013) and also report cut-off values.

In the present study, an app version of both the KSE-G and the IPO-16 was utilized using RShiny (Chang, Cheng, Allaire, Xie, & McPherson, 2017).

Procedure

Participants were sent an email with an outline of the experiment procedure and information regarding the place and date of their experiment session. At arrival on the experiment site, participants were provided detailed information about the type of data assessed in the experiment, the procedure of the assessment, their benefits in participating in the study, as well as contacts for further information and assurance that they could drop out of the experiment at any point in time. However, the underlying aim of the study was obscured in the information material and instead the study's aim was described as exploring the relationship between stress and personality, as well as physiological attributes. After receiving informed consent, the participants were asked to complete both the IPO-16 and KSE-G before undergoing the TSST-G as per protocol (Dawans et al., 2011). The only deviation from the standard protocol was the admission of only four participants at a time, compared to the six from the validation study (Dawans et al., 2011), as the premises did not allow for more participants at one timepoint. After the TSST-G, the ETQ was administered. Finally, the participants were debriefed about the aim of the study and compensated with 10€.

Ethics

The trial received ethical approval from the ethics committee of the Medical Faculty of the University of

Heidelberg, Germany (reference number: S-272/2017). The trial will be conducted in accordance with the European General Data Protection Regulation at all times. Participants will be identified by a study specific participant number during the experiment and anonymized at data aggregation. Names and any other identifying detail will not be included in any study data electronic file. In case sample sizes are very small (subgroups n>20), extra care will be taken by scaling the only personal variable, age, to mean 0 and standard deviation 1, to ensure that individual participants cannot be identified.

Data analysis

A priori estimation of the effect size between the statement categories for this study is not possible, as to our knowledge empirical data on the differences in certainty of retrospective assessments of statements of physiological, relational, and inner states is not available. Therefore, we chose to calculate power based on a medium effect of f^2=.25 between the categories, as a smaller effect could be the result of a flawed conceptualization of the paradigm. An *a priori* power analysis using GPower 3.1.9.2 (Faul, Erdfelder, Lang, & Buchner, 2007), using f^2=.25 as effect size, with an alpha of α=.05 and a power of β=.90, resulted in a sample size of n=54. Assuming a drop-out rate of 10% for participants withholding their data for analysis after debriefing, n=60 participants are to be recruited.

In the analysis of the primary hypothesis, the mean certainty ratings post-feedback per category are tested in a two-sided ANCOVA, controlling for gender and a major in psychology, since experience in psychological experiment design might undermine the relevancy aspect of the paradigm for psychology majors. Since all questionnaires utilized in the study are in app form, a forced answer format was chosen to achieve complete data for all participants with no missing values. R (version 3.4.1, R Development Core Team, 2008) is used in all statistical analyses.

Discussion and Conclusions

The described protocol for the validation of a new ET assessment aims to establish a comprehensive and theoretically grounded operationalization of ET in adults. Such new assessment method could provide support for a theory of personality disorder as a failure of communication between the individual and the social environment. It might also prove useful to measure ET pre- and post therapy to study probable predictors of therapeutic outcome. Additionally, being able to reliably measure ET might help disentangle ET, attachment, and mentalizing, three concepts that have historically been hard to separate because they tend to explain similar phenomena on a different level but are also closely related theoretically (*e.g.* Fonagy et al., 2015). Measuring all three constructs in one sample and mapping the relationships between them, ideally with an indicator of severity of personality disorder, ranging from normative to pathological, could provide a valuable empiric underpinning for future research in this field.

Despite these advantages, a number of potential limitations in our assessment need to be addressed. First, given the design of our procedure, its repetition may result in a loss of salience of the information provided and therefore in a lack of relevance. This is particularly unfortunate because repeating the procedure would be needed when attempting to apply it to the study of change, for example in psychotherapy research. In general, our procedure necessarily demands considerable time both from patients and therapist, which limits its applicability. Also, as there are no current alternative measures for ET it is difficult to externally validate the current paradigm except by using theoretically opposing constructs such as a diagnosis of Antisocial Personality Disorder or BPD with which ET should be negatively correlated.

However, if our paradigm will be successfully tested, it will provide the basis for designing more cost- and time-effective measures of ET. For example, a possible adaptation could investigate whether it can be operationalized without the stress inducing component (TSST-G), or whether the presence of a committee (but no job interview or arithmetic task) provides enough salience for the activation of ET. This could prove to be a viable step between an economically viable questionnaire but potentially limited validity and the very time consuming procedure outlined in this study. Another alternative would be to replace the rather rigorous TSST-G with a stressor such as the socially evaluated cold-pressor test (Minkley, Schröder, Wolf, & Kirchner, 2014; Schwabe, Haddad, & Schachinger, 2008). In this procedure, participants are exposed to a physical stressor, as they have to immerse their hand in ice water while they also are continuously observed and evaluated. This procedure could be adapted to include a more pronounced social evaluation aspect that makes it clear to the participants that the *expert* present during the experiment is evaluating them and to use this feedback akin to how the feedback from the committee is used in the present rendition of the ETA. Furthermore, this procedure could be adapted to further investigate the different types and role of ostensive cues in an adult population as well as to investigate the interaction with different psychopathologies.

References

Allison, E., & Fonagy, P. (2016). When is truth relevant? *The Psychoanalytic Quarterly*, *85*(2), 275-303. doi:10.1002/psaq.12074

Bateman, A., & Fonagy, P. (2016). *Mentalization Based Treatment for Personality Disorders* (2nd ed.). Oxford: Oxford University Press.

Beck, A. T., Davis, D. D., & Freeman, A. (2016). *Cognitive therapy of personality disorders* (3rd ed.). New York, London:

The Guilford Press.
Boesch, M., Sefidan, S., Ehlert, U., Annen, H., Wyss, T., Steptoe, A., & La Marca, R. (2014). Mood and autonomic responses to repeated exposure to the Trier Social Stress Test for Groups (TSST-G). *Psychoneuroendocrinology*, *43*, 41-51. doi:10.1016/j.psyneuen.2014.02.003
Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1491-1494. doi: 10.1037/0278-7393.27.6.1491
Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2017). Shiny: Web Application Framework for R. R package version 1.0.3. Available from https://CRAN.R-project.org/package=shiny
Corriveau, K., & Harris, P. L. (2009). Choosing your informant: weighing familiarity and recent accuracy. *Developmental Science*, *12*(3), 426-437. doi: 10.1111/j.1467-7687.2008.00792.x
Corriveau, K. H., Harris, P. L., Meins, E., Fernyhough, C., Arnott, B., Elliott, L., ... Rosnay, M. de. (2009). Young children's trust in their mother's claims: longitudinal links with attachment security in infancy. *Child Development*, *80*(3), 750-761. doi: 10.1111/j.1467-8624.2009.01295.x
Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148-153. doi: 10.1016/j.tics.2009.01.005
Daros, A. R., Uliaszek, A. A., & Ruocco, A. C. (2014). Perceptual biases in facial emotion recognition in borderline personality disorder. *Personality Disorders*, *5*(1), 79-87. doi: 10.1037/per0000056
Dawans, B. von, Kirschbaum, C., & Heinrichs, M. (2011). The Trier Social Stress Test for Groups (TSST-G): A new research tool for controlled simultaneous social stress exposure in a group format. *Psychoneuroendocrinology*, *36*(4), 514-522. doi: 10.1016/j.psyneuen.2010.08.004
Driskell, J. E., & Salas, E. (1991). Group decision making under stress. *Journal of Applied Psychology*, *76*(3), 473-478. doi: 10.1037/0021-9010.76.3.473
Egyed, K., Kiraly, I., & Gergely, G. (2013). Communicating shared knowledge in infancy. *Psychological Science*, *24*(7), 1348-1353. doi: 10.1177/0956797612471952
Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: attentional control theory. *Emotion (Washington, D.C.)*, *7*(2), 336-353. doi: 10.1037/1528-3542.7.2.336
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. doi: 10.3758/BF03193146
Foley, P., & Kirschbaum, C. (2010). Human hypothalamus-pituitary-adrenal axis responses to acute psychosocial stress in laboratory settings. *Neuroscience and Biobehavioral Reviews*, *35*(1), 91-96. doi: 10.1016/j.neubiorev.2010.01.010
Fonagy, P. (2004). *Affect regulation, mentalization, and the development of the self*. New York NY: Other Press.
Fonagy, P., & Allison, E. (2014). The role of mentalizing and epistemic trust in the therapeutic relationship. *Psychotherapy (Chicago, Ill.)*, *51*(3), 372-380. doi: 10.1037/a0036505
Fonagy, P., Luyten, P., & Allison, E. (2015). Epistemic petrification and the restoration of epistemic trust: a new conceptualization of borderline personality disorder and its psychosocial treatment. *Journal of Personality Disorders*, *29*(5), 575-609. doi: 10.1521/pedi.2015.29.5.575
Fonagy, P., Luyten, P., Allison, E., & Campbell, C. (2017). What we have changed our minds about: Part 2. Borderline personality disorder, epistemic trust and the developmental significance of social communication. *Borderline Personality Disorder and Emotion Dysregulation*, *4*, 9. doi: 10.1186/s40479-017-0062-8
Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, *59*(4), 601-613. doi: 10.1037/0022-3514.59.4.601
Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, *65*(2), 221-233. doi: 10.1037/0022-3514.65.2.221
Hagá, S., & Olson, K. R. (2017). Knowing-it-all but still learning: Perceptions of one's own knowledge and belief revision. *Developmental Psychology*, *53*(12), 2319-2332. doi: 10.1037/dev0000433
Harris, P. L., & Corriveau, K. H. (2011). Young children's selective trust in informants. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *366*(1567), 1179-1187. doi: 10.1098/rstb.2010.0321
Holmes, J., & Slade, A. (2017). Attachment in therapeutic practice. Sage.
Kemper, C. J., Beierlein, C., Bensch, D., Kovaleva, A., & Rammstedt, B. (2012). *Eine Kurzskala zur Erfassung des Gamma-Faktors sozial erwünschten Antwortverhaltens: die Kurzskala Soziale Erwünschtheit-Gamma (KSE-G)*. Available from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-339589
Kirschbaum, C., Kudielka, B. M., Gaab, J., Schommer, N. C., & Hellhammer, D. H. (1999). Impact of gender, menstrual cycle phase, and oral contraceptives on the activity of the hypothalamus-pituitary-adrenal axis. *Psychosomatic Medicine*, *61*(2), 154-162. doi: 10.1097/00006842-199903000-00006
Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The 'Trier Social Stress Test' - A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology, 28* (1-2), 76-81.
Leder, J., Hausser, J. A., & Mojzisch, A. (2013). Stress and strategic decision-making in the beauty contest game. *Psychoneuroendocrinology*, *38*(9), 1503-1511. doi: 10.1016/j.psyneuen.2012.12.016
Luyten, P. (2017). *Epistemic trust and BPD: An experimental approach.* Heidelberg, DE: ISSPD.
Matzke, B., Herpertz, S. C., Berger, C., Fleischer, M., & Domes, G. (2014). Facial reactions during emotion recognition in borderline personality disorder: a facial electromyography study. *Psychopathology*, *47*(2), 101-110. doi: 10.1159/000351122
Metcalfe, J., & Finn, B. (2011). People's hypercorrection of high-confidence errors: did they know it all along? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(2), 437-448. doi: 10.1037/a0021962
Metcalfe, J., & Finn, B. (2012). Hypercorrection of high confidence errors in children. *Learning and Instruction*, *22*(4), 253-261. doi: 10.1016/j.learninstruc.2011.10.004
Minkley, N., Schröder, T. P., Wolf, O. T., & Kirchner, W. H. (2014). The socially evaluated cold-pressor test (SECPT) for groups: Effects of repeated administration of a combined physiological and psychological stressor. *Psychoneuroendocrinology*, *45*, 119-127. doi: 10.1016/j.psyneuen.2014.03.022
Nicol, K., Pope, M., Sprengelmeyer, R., Young, A. W., & Hall, J. (2013). Social judgement in borderline personality disorder. *PloS One*, *8*(11), e73440. doi: 10.1371/journal.pone.0073440

Nolte, T. (2017). *Epistemic trust in adolescents and BPD patients: Two experimental approximations.* Heidelberg, DE: ISSPD.

Paulhus, D. L. (2015). Socially desirable responding: The evolution of a construct. In H. Braun, Jackson, D.N., & D. E. Wiley (Eds.), *The Role of Constructs in Psychological and Educational Measurement* (pp 49-69). London: Routledge.

R Development Core Team. (2008). *A language and environment for statistical computing*. Vienna, Austria.

Schnell, K., & Herpertz, S. C. (2018). Emotion regulation and social cognition as functional targets of mechanism-based psychotherapy in major depression with comorbid personality pathology. *Journal of Personality Disorders*, *32*(Supplement), 12-35. doi: 10.1521/pedi.2018.32.supp.12

Schwabe, L., Haddad, L., & Schachinger, H. (2008). HPA axis activation by a socially evaluated cold-pressor test. *Psychoneuroendocrinology*, *33*(6), 890-895. doi: 10.1016/j.psyneuen.2008.03.001

Sharp, C., Ha, C., Carbone, C., Kim, S., Perry, K., Williams, L., & Fonagy, P. (2013). Hypermentalizing in adolescent inpatients: treatment effects and association with borderline traits. *Journal of Personality Disorders*, *27*(1), 3-18. doi: 10.1521/pedi.2013.27.1.3

Sharp, C., Pane, H., Ha, C., Venta, A., Patel, A. B., Sturek, J., & Fonagy, P. (2011). Theory of mind and emotion regulation difficulties in adolescents with borderline traits. *Journal of the American Academy of Child and Adolescent Psychiatry*, *50*(6), 563. doi: 10.1016/j.jaac.2011.01.017

Sperber, D., Clement, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359-393. doi: 10.1111/j.1468-0017.2010.01394.x

Sperber, D., & Wilson, D. (1995). Relevance. Communication and cognition. Oxford, Cambridge, MA: Blackwell Publishers Ltd.

Starcke, K., & Brand, M. (2012). Decision making under stress: a selective review. *Neuroscience and Biobehavioral Reviews*, *36*(4), 1228-1248. doi: 10.1016/j.neubiorev.2012.02.003

Wilson, D., & Sperber, D. (2012). *Meaning and relevance*. Cambridge: Cambridge University Press.

Zimmermann, J., Benecke, C., Hörz, S., Rentrop, M., Peham, D., Bock, A., ... Dammann, G. (2013). Validierung einer deutschsprachigen 16-Item-Version des Inventars der Persönlichkeitsorganisation (IPO-16). *Diagnostica*, *59*(1), 3-16. doi: 10.1026/0012-1924/a000076