Preliminary Psychometric Scale Development Using the Mixed Methods Delphi Technique

Yavor Dragostinov

University of Edinburgh

Daney Harðardóttir

King's College London

Peter Edward McKenna

Heriot-Watt University

David A. Robb

Heriot-Watt University

Birthe Nesset

Heriot-Watt University

Muneeb Imtiaz Ahmad

Swansea University

Marta Romeo

University of Manchester

Mei Yii Lim

Heriot-Watt University

Chuang Yu

University of Manchester

Youngkyoon Jang

Imperial College London

Mohammed Diab

Imperial College London

Angelo Cangelosi

University of Manchester

Yiannis Demiris

Imperial College London

Helen Hastie

Heriot-Watt University

Gnanathusharan Rajendran

Heriot-Watt University

Correspondence concerning this article should be addressed to Yavor Dragostinov Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK. E-mail: y.dragostinov@sms.ed.ac.uk

This manuscript has been accepted for publication in *Methods in Psychology*

Abstract

This study implemented a Delphi Method; a systematic technique which relies on a panel of experts to achieve consensus, to evaluate which questionnaire items would be the most relevant for developing a new Propensity to Trust scale. Following an initial research team moderation phase, two surveys were administered to academic lecturers, professors and Ph.D. candidates specialising in the fields of either individual differences, human-robot interaction, or occupational psychology. Results from 28 experts produced 33 final questionnaire items that were deemed relevant for evaluating trust. We discuss the importance of content validity when implementing scales, while emphasising the need for more documented scale development processes in psychology. Furthermore, we propose that the Delphi technique could be utilised as an effective and economical method for achieving content validity, while also providing greater scale creation transparency.

Keywords: Delphi Method, Scale Development, Propensity to Trust, Mixed Methods, Instrument Development

1. Introduction

When developing a questionnaire, expert evaluation of items is considered an important step in establishing content validity (Boateng et al., 2018), making it an essential part of scale development and validation. This occurs after the careful identification of knowledgeable experts from the domain of interest and/or with experience in scale development (DeVellis, 2012; Morgado et al., 2017). After a pool of items is generated, experts evaluate each item to determine their relevance to the domain of interest. In psychological assessment, the number of expert judges usually ranges from 5 to 7 (Haynes et al., 1995). Boateng and colleagues (2018, p.7) stated that ‘an increase in the number of experts has been found to increase the robustness of the findings (Haynes et al., 1995; Lynn, 1986)’. A method that assesses content validity using multiple expert judges in a systematic way is the *Delphi Method* – an instrument of gathering consensus that has been relatively unexplored, especially in the field of personality psychology (Iqbal & Pipon-Young, 2009).

One concept that would benefit from a Delphi Method is trust, in particular propensity to trust.[1] While trust has been shown to be very important in multiple social dynamics (Ferguson & Peterson, 2015), as Patent and Searle (2019) state, “there has been a surprisingly lack of attention to understand why and how some individuals trust more readily than others (Colquitt, Scott, & LePine, 2007; Dietz, 2011)” (p. 136).

In this paper, we will implement a Delphi Method as a stepping-stone in the psychometric scale development of a Propensity to Trust scale. Adopting the Delphi method will provide an innovative approach to scale development, as the method involves both quantitative and qualitative data, and makes the process of content validity more transparent. Furthermore, we will show how the Delphi Method can be used to garner expertise from an interdisciplinary field (e.g., researchers of artificial intelligence and trust) and narrow down a

[1] More information on the larger Propensity to Trust scale project is available at https://trust.tas.ac.uk/.

large number of related items from the literature to a set of core items with high expert consensus.

1.1. The Delphi Method

The first Delphi Method study was implemented by researchers at the RAND Corporation (Santa Monica, California) as part of a military defence project. A documented proposal to use the Delphi Method for non-military purposes also came from the RAND Corporation (Helmer & Quade, 1963).

The Delphi Method is a survey item refinement process based on two or more rounds. In each round, participants are asked the rate and reflect on the relevance of survey items to a chosen topic. As Barrett and Heale (2020) wrote, "The questions for each round are based in part on the findings of the previous one, allowing the study to evolve over time in response to earlier findings" (p. 68). Participants are presented with the results from previous rounds – including their own responses allowing them to reflect on the views of others and, possibly, reposition their own evaluation accordingly (Keeney et al., 2006). This provides the opportunity for experts to offer feedback regarding the strengths and weaknesses of other responses (Barrett & Heale, 2020). To finish, the findings of each round are shared with the group anonymously, meaning that their answers go directly to the group coordinator. This helps to avoid any bias that could result from the influence of the personality or status of other participating experts (Landeta, 2006; Barrett & Heale, 2020).

Within the Delphi Method, there can be variation for how many rounds there are, how many participants are required, and how consensus is determined. For instance, the number of experts will vary depending on the topic, as well as the time and resources available to the researcher (Iqbal & Pipon-Young, 2009). Hsu and Sandford (2007) recommend 10 to 15 participants, while Turoff (2002) recommends anywhere between 10 and 50.

Three or more rounds are preferable for studies measuring consensus, while fewer rounds are recommended for studies measuring opinions (Iqbal & Pipon-Young, 2009). Rounds of the Delphi Method may also continue until a consensus is reached. However, such an approach could compromise response rate and enthusiasm. For this reason, three rounds tend to be sufficient (Stone Fish & Busby, 2005).

There is no agreed upon gold standard for achieving consensus in Delphi studies. However, studies have used a recommended criteria of 70% as a proxy for agreement and 15% for disagreement (Fackrell et al., 2017; Williamson et al., 2012). For example, if the first round of a Delphi study shows that 70% of participants deem an item as relevant, while no more 15% of them deem it as irrelevant, this would indicate sufficient consensus to keep the item for subsequent rounds.

When evaluating content validity, the approach assumes that you have an accurate description of the domain of interest – something that is not always true (Trochim & Donelly, 2006). Content validity is still a subjective form of measurement because it relies on individuals' perception that would otherwise be difficult to measure (Hufford, 2021). However, what separates the Delphi Method approach beyond other methods of item selection is the inclusion of outside expertise that would not otherwise be present. Furthermore, the Delphi Method has demonstrated to be an excellent tool in establishing content validity (Colton & Hatcher, 2004) – potentially leading to a clearer theoretical factor structure that will be assessed at a later stage by a factor analysis.

Expert selection criteria are crucial to successful use of the Delphi method. The Delphi Method could be compromised if the panellists lack the relevant knowledge and qualifications (Keeney et al., 2001). Additionally, including experts from different disciplines, who can offer a unique perspective, is recommended for the development of a credible questionnaire (Linstone & Turoff, 2002; Iqbal & Pipon-Young, 2009). Moreover, a mixture of open-ended and Likert scale questions is considered an appropriate way of conducting a Delphi Method study (Iqbal & Pipon-Young, 2009; Fackrell et al., 2017), as it allows for both consensus evaluation and brainstorming.

The Delphi Method uses a mixed-methods (MM) approach, incorporating both quantitative and qualitative methods, attaining a more complete picture of the subject matter (Iqbal & Pipon-Young, 2009). A properly planned MM design can lead to complementary findings between each data type (Hughes, 2016). For example, the MM approach allows researchers to identify items that are perceived as relevant (through quantitative feedback), whilst recognising the need to improve their wording (as per the qualitative feedback). This process strengthens the validity of the item selection and refinement process, as decisions about item inclusion and item amendments are informed by a combination of numeric assessment and qualitative feedback. This is especially pertinent for items that may be considered borderline by a research team, but through the Delphi Method, achieve below threshold consensus from examination of the numeric evaluations and accompanying qualitative feedback. Through this process, the Delphi Method reduces the risk of bias in survey generation and refinement.

1.2. Propensity to Trust Measurement

Frazier et al. (2013) describes propensity to trust (P2T) as ‘a general willingness to trust others, regardless of social and relationship-specific information’ (p. 80). Even though P2T has been considered highly relevant for the research of trust in organisations, there has been little attention paid to the quality of the existing scales used to measure this construct. As Patent and Searle (2019) state, ‘several authors have expressed concern about the lack of reliable measures, to support their development of new measures (Ashleigh, Higgs, & Dulewicz, 2012; Frazier et al., 2013; Schoorman, Mayer, & Davis, 2007)’ (p. 137). In 2019, Patent and Searle conducted a qualitative meta-analysis that evaluated the P2T literature

(1966-2018) and identified 26 measures from 179 studies. The authors assessment of these scales revealed substantial methodological concerns and considerable variety in quality.

As the predominant view is that trust is a stable personality trait (Evans & Ravelle, 2008), most quantitative studies have relied on evaluating the psychometric properties of the construct using surveys. In contrast, qualitative studies that have measured trust have often taken a clinical approach involving semi-structured interviews – e.g., doctor-patient relationships (Dang et al., 2017; Wright et al., 2004), therapeutic relationship online (Fletcher-Tomenius & Vossler, 2009) and the complexities of informed consent to clinical trials (Rost et al., 2021).

1.3. Relevance to Psychological Research

While the Delphi Method approach has been used in psychological research (Graham & Milne, 2003; Haggard & Haste, 1986; Haste et al., 2001; Jeffery et al., 2000; Petry et al., 2007), we are not aware of it being used for scale development prior to this work. This is surprising, as the Delphi Method is suitable for establishing content validity; i.e., assessing how accurately items map on to the concept they are presumed to measure (DeVellis, 2012).

Lastly, there is a tendency in the measurement literature to create and use scales with no evidence of their systematic development (Flake & Fried, 2020). As Flake and Fried report, "This has been documented in literature on emotions (69% of scales sampled; Weidman et al., 2017), education and behaviour (40–90% of articles sampled; Barry et al., 2014), and social psychology and personality (40% of scales sampled; Flake et al., 2017)." (p. 463). Moreover, measurement details are frequently unreported in public psychological literature; out of 433 scales that were reviewed from a random sample of studies published in a popular psychological journal in 2014, 40% were reported without information about their source, 19% were reported without the number of items, and 9% were reported without the response scale (Flake et al., 2017). Applying the Delphi Method to scale development will prevent such issues, while also detailing the scale development process.

2. Method

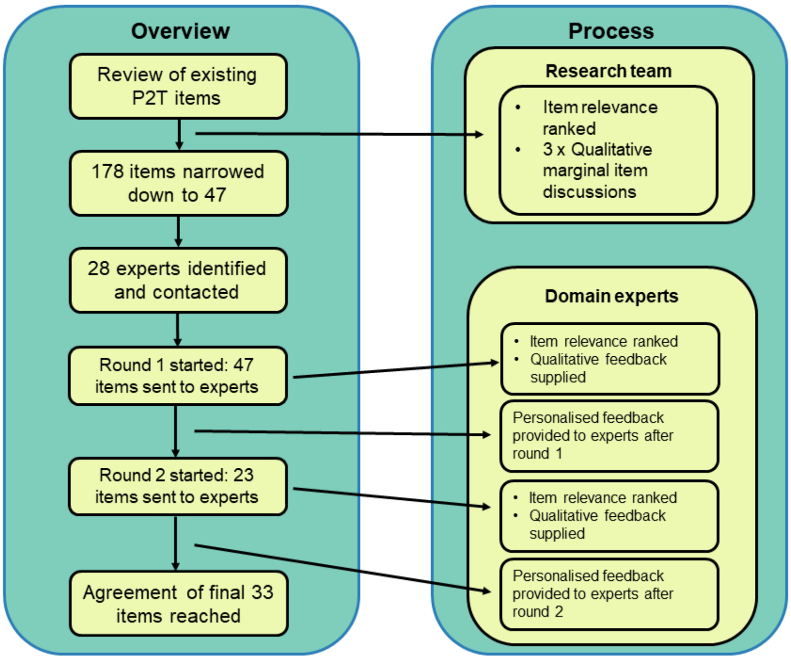
2.1. Participants

Experts were targeted in the fields of differential psychology, occupational psychology, human-robot interaction (HRI), and trust research. To be eligible, participants had to be English speaking, hold or be working towards a Ph.D., and have expertise in one of the following domains: trust, psychometrics, individual differences, occupational psychology, and human-robot interaction. As an incentive, participants were informed they would be recognised in the Acknowledgement section of publications generated from the data.

A total of 28 participants that completed the initial round were invited to participate in consecutive rounds – drop outs (*n* = 5) meant attrition between rounds was 18%. Mean age in round 1 was 37.98 years (SD = 9.76), in round 2 mean age was 37.81 years (SD = 9.71). A breakdown of further participant characteristics can be found in Table 1.

2.2. Procedure

Figure 1. Flow diagram of Delphi Method processes.


Overview
Review of existing P2T items
178 items narrowed down to 47
28 experts identified and contacted
Round 1 started: 47 items sent to experts
Round 2 started: 23 items sent to experts
Agreement of final 33 items reached
Process
Research team
• Item relevance ranked
• 3 x Qualitative marginal item discussions
Domain experts
• Item relevance ranked
• Qualitative feedback supplied
Personalised feedback provided to experts after round 1
• Item relevance ranked
• Qualitative feedback supplied
Personalised feedback provided to experts after round 2

The design and methodology of this study was preregistered[1]. To identify potentially relevant indicators of P2T to include in this Delphi survey, existing tools measuring trust in different contexts were identified[1] (Rosenberg, 1956; Wrightsman, 1964; Rotter, 1967; MacDonald et al., 1972; Schuessler, 1982; Costa & McCrae, 1985; Yamagishi, 1986; Schoorman et al., 1996; Jarvenpaa et al., 1998; Goldberg, 1999; Gefen, 2000; Lee & Turban, 2001; McKnight et al., 2002; Hu & Kelley (2003); Valenzuela et al., 2008; Evans & Revelle ,2008; Kantsperger & Kunz, 2010; Ashleigh et al., 2012; Frazier et al., 2013) and a list of existing items was collated (item N = 178). In the first round of iteration, six researchers (including experts from the domains of individual differences, psychometrics, HRI) independently ranked the 178 items rating each according to one of three acceptance categories: Yes (keep); No (omit); Maybe (discuss at moderation). This initial screening was undertaken to optimise the time and effort of the Delphi respondents in two ways: to exclude obviously irrelevant items from the item pool; and to reduce the size of the item pool so that the demands of the Delphi survey were appropriate. Reasons for item exclusion at this initial

[1] The preregistration can be found https://osf.io/hdkra.
[1] Serendipitously, a qualitative meta-analysis by Patent and Searle (2019) listed all the existing Trust measures.

screening were as follows: item measured trust instead of P2T (e.g., ‘I trust others.’); item was too context specific (e.g., ‘Most elected public officials are really sincere in their campaign promises.’); item was too simple (e.g., ‘I believe that people seldom tell you the whole story.’); item was too verbose (e.g., ‘I am usually suspicious of people until I have had plenty of time to get to know them and know they can be trusted.’; item was awkwardly constructed (e.g., ‘Employees will not work hard or do quality work unless managers closely monitor their work.’); item focused on the broader concept of morality (e.g., ‘I believe that people are basically moral.’); item was irrelevant to P2T (e.g., ‘Most professionals are very knowledgeable in their chosen field.’). Numeric values were then assigned to each of these categories (Keep = 1, No = 0, Maybe = 0.5) and a score was generated by summating all of the research team’s rankings per item. So, the maximum item score was 6 and the minimum 0. These summated item scores were then rank ordered by numeric superiority so that items with high acceptance were at the top of the list and those with low acceptance at the bottom. Following these numeric evaluations, the six researchers held three weekly moderation meetings to discuss items with borderline numeric rankings (e.g., item score = 2) and item theme relevance. A colour coding scheme of Green = ‘Definitely keep’, Orange = ‘Not sure’, Red = ‘Definitely discard’ was used at this stage to focus discussions on ambiguous items and item relevance to P2T. Following discussion of definite exclusions, and ambiguous or borderline items, the list was refined from 178 down to 47. Some small amendments to the items were then made based on researcher feedback, including adding a pronoun to items (e.g., ‘Suspect hidden motives in others.’ became ‘I suspect hidden motives in others.’) and minor grammatical changes. The final 47 items were then sent out to expert participants in electronic survey form to begin the Delphi Method.

Participants were identified by the research team based on their status as subject matter experts.[1] Additional participants were also recruited through researcher networks (e.g., HRI Trust Slack channel) and a general call to interested parties. Identified participants were then approached via email and sent a link to the first Delphi round.[1]

For our online Delphi Method participants were asked to rate each item on a five-point Likert-scale indicating the relevance of the item as a measure of P2T, ranging from 'very irrelevant' to 'very relevant'. Participants were given the following P2T definition adapted from Frazier et al., (2013) to standardise participants P2T understanding, 'A general willingness to trust other people, regardless of social and relationship-specific information'.[1]

Qualitative data was collected using open-ended text responses. Following every page (10 items were presented per page), participants were provided with a text entry space in which they were invited to provide feedback regarding the respective 10 items that were shown on the page. A final text-entry space appeared at the end of the survey which asked the experts to provide overall feedback regarding the whole measurement. To prevent participation fatigue, all open-ended text spaces were made optional. After every round, the research team got together and discussed the qualitative feedback of every expert in accordance with the respective page of the survey.[1] Expert qualitative feedback was used to guide decisions on item inclusion, as well as to refine items in future rounds. Qualitative feedback was also used to refine items between rounds. Unchanged items with strong consensus (based on the criteria defined in the Analysis section) were excluded from the second round. The second round was sent out to participants approximately two and a half months after. The experts were asked to

[1] The names of all the experts can be found in the 'Acknowledgements' section.

[1] The qualitative feedback from experts can be found in the online repository.

[1] We modified Frazier et al.'s (2013) definition 'a general willingness to trust others, regardless of social and relationship-specific information' (p. 80) to examine only human interpersonal trust, so 'others' was changed to 'other people'.

[1] The quantitative feedback refers to the score each expert gave to each item, respectively. The qualitative feedback refers to the optional open-ended text box in which experts were invited to write what they think overall for every set of 10 items (one per page).

repeat the process of rating the relevance of items as indicators of P2T and provide open format feedback. Apart from the survey link, the experts were sent personalised tables with all their individual responses to each item, as well as the overall mean and median of all the participants[1]. After the second round the Delphi Method was terminated. Items that had not reached consensus post-Delphi were further moderated by the research team, and a decision made to include them in the final list or not.

Both survey rounds were carried out using the online survey platform Qualtrics. Both the qualitative and quantitative data were analysed to determine the items for the final P2T measure.

2.3. Analysis

All quantitative data analysis was performed in R version 3.6.1 (R Studio Team, 2018). Item-level survey data was analysed, both within each round and across the repeated rounds. This was done using descriptive statistics, item means and medians as well as measures of variance for each round. Stability between rounds was explored through changes in mean difference ratings of items, subtracting the mean at timepoint 1 from the mean at timepoint 2.

The main outcome of this study is expert consensus on the relevance of items as indicators of P2T. Because there is no agreed upon gold standard for achieving consensus in Delphi studies, we used the following criteria because they are simple and frequently used in the literature (e.g., Fackrell et al., 2017; Williamson et al., 2012). To make inferences about consensus for individual items the percentage of participants that chose each response option was calculated. For analysis purposes the response options ‘relevant’ and ‘very relevant’ were collapsed into one category referred to as ‘relevant’. Likewise, ‘very irrelevant’ and ‘irrelevant’ were collapsed into a category labelled ‘irrelevant’. Consensus that an item is

[1] An example of a mock personalized table can be found in the online repository.

relevant was reached if $\geq$ 70% of participants deemed an item to be relevant, while no more than 15% judged the same item irrelevant. Alternatively, consensus that an item is irrelevant to the construct of P2T was defined as at least 70% of participants judging the item as irrelevant and no more than 15% judging it relevant. Open-ended text responses were analysed thematically.

2.4. Data Availability Statement

Data, materials, and reproducible code can be found at https://osf.io/hdkra/.

3. Results

3.1. Round 1

In round one, 25 out of 47 items reached consensus according to the above criteria. All 25 items were deemed to be relevant by at least 70% of participants and irrelevant by less than 15% (see Table 2). The other 22 items that did not reach consensus were automatically included in the second round of the Delphi, along with any items that were refined or amended based on qualitative feedback to open format questions. A total of 7 items were refined between rounds 1 and 2 (for items that were reworded and the reasons to do so, please see Table 3).).

3.2. Round 2

In round 2, 26 items were rated by the subject matter experts. A total of 6 (23.1%) items reached consensus and were deemed relevant (see Table 4). The research team held a moderation discussion to decide which of the remaining 20 items would be retained. A further 6 items that did not reach consensus were included. Reasons for these items' inclusion were that they were borderline consensus items (e.g., 68% relevant) and that generally, a conservative approach to exclusion was adopted due to the small sample size. Additionally, some items related to distrust were retained – despite not reaching consensus – as this sub-domain of P2T was not well represented in the retained items.

4. Discussion

In this study, we conducted a two-round Delphi Method to assess the content validity of items that are meant to measure P2T. There were two main reasons this work was conducted: 1) To establish a reliable foundation in developing a new Propensity to Trust Scale. 2) To utilise the Delphi Method, a particularly underused tool in psychometrics, as an important step in scale development.

One of the main reflections from using the Delphi Method was that empowering field experts to give qualitative feedback was crucial to item selection, and we strongly recommend researchers endeavouring to develop a scale from previous items to do the same. We adjusted multiple items after changing the wording of them following the Round 1 qualitative feedback (see Table 3). If we had relied solely on quantitative data alone, items which contain valuable information would have likely been eliminated. Furthermore, the qualitative data provided valuable insight in distinguishing trust and distrust; items that assumed honesty and openness are core parts of people's understanding of trust; items that described faith in human nature were dependent on one's individual beliefs about human nature and more. Most experts provided elaborate reasons and/or references for why they thought certain items were not appropriate. This not only made our decision-making easier, but also provided useful information regarding trust and psychometrics, which ultimately improved our understanding of the topic. Lastly, the qualitative approach encompasses a wider range of viewpoints, research methods, and interpretations of human experiences (Denzin & Lincoln, 2002). This is particularly relevant when you are interacting with experts from different fields such as those in our survey. Providing experts with the freedom to express their thought processes regarding the subject matter enhanced the internal discussion within our research group during the interpretation of results.

The increased awareness and emphasis on the replication crisis in psychology has been extremely important for psychological science. We support these concerns, while we also highlight the need for more rigorous systematic development and transparency when creating scales. This has been illustrated appropriately by Flake and Fried (2020), when they discuss the concept of depression as an example: ‘even if the sample size of depression trials is increased, studies are adequately powered, analytic strategies are preregistered, and *p*-hacking stops, researchers can still be left wondering if they were ever measuring depression at all.’ (p. 464).

Furthermore, the Delphi Method is particularly economical in terms of finances and participant time. For example, in the current work, participants were not financially reimbursed for their contribution. This lack of resource intensity might differ based on the type of expertise a researcher is seeking when conducting a Delphi Method. However, in our experience of recruiting experts for this project, academics (especially from psychological fields), appear to be open and willing to participate in multiple stages of assessment of a measurement tool, even when they did not personally know the research team that is conducting the study. This implies that researchers, junior or established, can conduct a Delphi Method with relative ease by drawing on their network of academic peers.

In applying the Delphi Method, scale developers can not only mitigate the risk of ‘group think’ within their own research team (so contaminating or biasing their process), but they can also act positively by involving a much wider community of contributors. Thus, promoting more diverse judgements on which to draw upon. As the equal application of criteria is intrinsic to the process, equal weight is given to every gathered judgement helping to ensure that equality is combined with diversity in producing the final output.

4.1. Future Direction, including broader usage and appeal of Propensity to Trust

As robots and autonomous systems become increasingly present in our lives, we need reliable measurement tools to study dispositional human factors. To achieve this, we will need a reliable and valid measure of P2T. Trust is just one of the many constructs that are becoming increasingly important in HRI (Hancock et al. 2020). For example, as we tease out individual factors (e.g., user expertise, technology affinity) from situational factors (e.g., the type of robot platform being used, the task we are asking it to do), we need to work out the extent to which a variation in trust can be attributed to individual differences in P2T, or the robot itself and the task the robot has been set. Further, Ullman et al. (2021) have counselled the field of HRI falling into the same replication crisis that has beleaguered psychology. So, in this paper, we provide a process for HRI researchers to follow when creating their own scales.

With the development and integration of natural language processing into the trait measurement literature, it would be interesting to entertain the idea of using a sentiment analysis for text and qualitative data (Alexander & Hudson, 2022) alongside the analysis of each Delphi round. That could provide an even more rigorous assessment of content validation.

Presently, the UKRI TAS Node in Trust[1] is investigating dispositional factors that affect people's likelihood to trust a robot, including attitudes towards robots and P2T. While there are fairly robust and well adopted tools to measure attitudes towards robots (e.g., the Negative Attitudes Towards Robots Scale; Nomura et al., 2006), fewer efforts have been made to develop a similarly robust tool for the measurement of P2T. So, the TAS Node in Trust team plan to use the findings of the Delphi study to create a more empirically sound measure of P2T and deploy this new tool to examine how P2T varies the likelihood that people will trust a robot.

[1] https://trust.tas.ac.uk/

We want to emphasize that the Delphi Method is a starting point to a long process of psychometric stages to ensure that a measurement is sufficient. We intend to conduct exploratory and confirmatory factor analyses, construct and discriminatory validity checks and test-retest estimates in future work.

There are some limitations to the Delphi method that are worth consideration. As Barrette and Heale (2020) point out, relative to more conventional item selection processes (e.g., research team moderation), the Delphi method is time consuming, which can result in high drop out rates between rounds. In the present research, only 5 experts dropped out between rounds 1 and 2. We suspect that we were able to maintain a good level of expert participation because of our pre-Delphi item moderation phase. During this phase, we reduced what was a long list of items down to something more manageable (from 178 to 47); especially for external persons volunteering their time in kind. Our team included academics from a variety of disciplines, including computer science and psychometrics, with different motivations to engage in the research. This blend of expertise and varying objectives helped to mitigate some bias in the moderation process, though intra-institutional biases may have been unavoidable (e.g., peers who share a similar perspective on P2T). So, those who wish to conduct a Delphi method in future should consider the size of the item pool for analysis, and whether a pre-Delphi moderation stage should be included to safeguard against between-round dropout. Another issue inherent in the Delphi process is the consensus criteria used to select items. In our study we used the boundaries set by Fackrell et al. (2017) and Williamson et al. (2012), whereby items with agreement consensus ≥ 70% were retained and 15% omitted, and vice-versa for disagreement consensus (e.g., consensus ≥ 70% were omitted and 15% retained). This is by no means an ideal assessment criterion, as one could argue that consensus of 70% indicates a 30% levels of disagreement. To overcome this, we used qualitative data to supplement item inclusion decisions. Future studies adopting the Delphi Method will lead to a larger pool of data which to establish item inclusion criteria.

5. Conclusion

In the present work, we deployed the Delphi method – a previously underused form of scale development – to create a valid and reliable P2T scale. To begin, a combination of qualitative and quantitative assessment was adopted by the research team, using a mix of moderation discussion and numeric item rankings. This initial selection process refined the list of items prior to conducting a two-round Delphi survey. Using statistically bound inclusion/exclusion criteria, expert rankings, and expert feedback the Delphi method iterated the list of items down further. With a few final modifications, what began as a list of 147 items was eventually narrowed down to 33. By adopting the Delphi Method, we are confident our final list of items is objectively relevant to the domain of P2T.

The Delphi method is a useful tool for establishing consensus for phenomena that are ill-defined or suffer from an overabundance of independently created surveys. It is especially helpful moving forward in psychology, HRI, and sciences investigating social phenomena, as improving the overall content validity of survey items will tackle ongoing issues with reproducibility.

Ethics Approval

This study was granted ethical approval from the University of Edinburgh on the 22nd of April 2021, reference number 233-2021/3.

Acknowledgements

This work was supported by the UKRI Node on Trust (EP/V026682/1). We would like to sincerely thank all the experts that took part in this study. Here are their names in alphabetical order:

Alice Diaz, Alistair Soutter, Amanda Ferguson, Beatrice Mahoney, Brett Israelsen, Daniel Farrelly, Drew Altschul, Esperanza Badaya, Friederike Eyssel, Jasmin Bernotat, Joléne Van Der Mescht, Lisanne de Moor, Martin Corley, Martin Porcheron, Martin Ross, Michael Baer, Michelle Luciano, Paul-Alan Armstrong, Peggy Wu, René Mõttus, Ross Stewart, Samuel Henry, Stuart Ritchie, Tom Booth, Volker Patent, Xiangling Hou, and Yuzhan Hang.

References

Alexander P C, Hudson G (2022). *transforEmotion: Sentiment analysis for text and qualitative data*. R package version 0.1.1.

Ashleigh, M. J., Higgs, M., & Dulewicz, V. (2012). A new propensity to trust scale and its relationship with individual well-being: Implications for HRM policies and practices. *Human Resource Management Journal, 22(*4), 360–376.

Barrett, D., & Heale, R. (2020). What are Delphi studies?. *Evidence-based nursing, 23*(3), 68-69.

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior, 41,* 12-18. doi:10.1177/1090198113483139

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in public health, 6*, 149.

Colquitt, J., Scott, B., & LePine, J. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology, 92*(4), 909–927.

Colton, S., & Hatcher, T. (2004). The Web-Based Delphi Research Technique as a Method for Content Validation in HRD and Adult Education Research. *Online Submission*.

Costa, P. T., & McCrae, R. R. (1985). *Neo PI-R professional manual.* Odessa, FL: Psychological Assessment Resources.

Dang, B. N., Westbrook, R. A., Njue, S. M., & Giordano, T. P. (2017). Building trust and rapport early in the new doctor-patient relationship: a longitudinal qualitative study. *BMC medical education*, *17*(1), 1-10.

Denzin, N. K., & Lincoln, Y. S. (2002). The qualitative inquiry reader. London: Sage Publications.

Devellis, R. (2012). *Scale Development Theory and Applications.* Sage Publications, New York.

Dietz, G. (2011). Going back to the source: Why do people trust each other? Journal of Trust Research, 1(2), 215–222.

Evans, A. M., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality, 42*(6), 1585–1593. https://doi.org/10.1016/j.jrp.2008.07.011

Evans, A., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. Journal of Research in Personality, 42(6), 1585–1593.

Fackrell, K., Smith, H., Colley, V., Thacker, B., Horobin, A., Haider, H. F., ... & Hall, D. A. (2017). Core outcome domains for early phase clinical trials of sound-, psychology-, and pharmacology-based interventions to manage chronic subjective tinnitus in adults: the COMIT'ID study protocol for using a Delphi process and face-to-face meetings to establish consensus. *Trials*, *18*(1), 1-11.

Ferguson, A. J., & Peterson, R. S. (2015). Sinking slowly: Diversity in propensity to trust predicts downward trust spirals in small groups. *Journal of Applied Psychology*, *100*(4), 1012.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456-465.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research. *Social Psychological and Personality Science, 8,* 370-378. doi:10.1177/1948550617693063

Fletcher-Tomenius, L., & Vossler, A. (2009). Trust in online therapeutic relationships: The therapist's experience. *Counselling Psychology Review*, *24*(2), 24-34.

Frazier, M. L., Johnson, P. D., & Fainshmidt, S. (2013). Development and validation of a propensity to trust scale. *Journal of Trust Research, 3*(2), 76–97.

Gefen, D. (2000). *E-commerce: The role of familiarity and trust.* Omega, 28(6), 725.

Goldberg, L. R. (1999). *A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models.* In I. Mervielde, I. Deary, F.

De Fruyt, & F. Ostendorf (Eds.), Personality psychology in Europe (Vol. 7, pp. 7–28). Tilburg: Tilburg University Press.

Graham, L. & Milne, D. (2003). Developing basic training programmes: A case study illustration using the Delphi method in clinical psychology. *Clinical Psychology and Psychotherapy, 10*, 55–63.

Haggard, M. & Haste, H. (1986). One generation after 1984: Psychology in the year 2010. *Bulletin of the British Psychological Society, 39*, 321–324.

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2021). Evolving trust in robots: specification through sequential and comparative meta-analyses. *Human factors, 63*(7), 1196-1229.

Haste, H., Hogan, A. & Zacharious, Y. (2001). Back (again) to the future. *The Psychologist, 14*(1), 30–33.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological assessment*, *7*(3), 238.

Helmer & Quade, E. (1963). An approach to the study of developing economy by operational gaming. (Publication No.2718). Santa Monica, CA: Rand Corporation.

Hsu, C. C., & Sandford, B. A. (2007). The Delphi technique: making sense of consensus. *Practical assessment, research, and evaluation*, *12*(1), 10.

Huff, L., & Kelley, L. (2005). Is collectivism a liability? The impact of culture on organizational trust and customer orientation: A seven-nation study. *Journal of Business Research, 58*(1), 96–102.

Hufford, B. (2022) *The 4 types of validity in research design (+3 more to consider), ActiveCampaign*. Available at: https://www.activecampaign.com/blog/validity-in-research-design (Accessed: November 9, 2022).

Iqbal, S. and Pipon-Young, L., 2009. *The Delphi method | The Psychologist*. [online] Thepsychologist.bps.org.uk. Available at: <https://thepsychologist.bps.org.uk/volume-22/edition-7/delphi-method> [Accessed 28 January 2022].

Jarvenpaa, S. L., Knoll, K., & Leidner, D. E. (1998). Is anybody out there? Antecedents of trust in global virtual teams. Journal of Management *Information Systems, 14*(4), 29–64.

Jeffery, D., Ley, A., Bennun, I. & McLaren, S. (2000). Delphi survey of opinion on interventions, service principles and service organisation for severe mental illness and substance misuse problems. *Journal of Mental Health, 9*, 371–384.

Kantsperger, R., & Kunz, W. H. (2010). Consumer trust in service companies: A multiple mediating analysis. *Managing Service Quality: An International Journal, 20*(1), 4–25.

Keeney, S., Hasson, F., & McKenna, H. (2006). Consulting the oracle: ten lessons from using the Delphi technique in nursing research. *Journal of advanced nursing*, *53*(2), 205-212.

Landeta, J. (2006). Current validity of the Delphi method in social sciences. *Technological forecasting and social change*, *73*(5), 467-482.

Lee, M. K., & Turban, E. (2001). A trust model for consumer internet shopping. *International Journal of Electronic Commerce, 6*(1), 75–91.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing research*.

MacDonald, A. P., Kessel, V. S., & Fuller, J. B. (1972). Self-disclosure and two kinds of trust. *Psychological Reports, 30*(1), 143–148.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for ecommerce: An integrative typology. *Information Systems Research, 13*(3), 334–359.

Morgado, F. F., Meireles, J. F., Neves, C. M., Amaral, A., & Ferreira, M. E. (2017). Scale development: ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica*, *30*.

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006). Measurement of negative attitudes toward robots. *Interaction Studies*, *7*(3), 437-454.

Petry, K., Maes, B. & Vlaskamp, C. (2007). Operationalizing quality of life for people with profound multiple disabilities: A Delphi study. *Journal of Intellectual Disability Research, 51*(1), 334–349.

Rosenberg, M. (1956). Misanthropy and political ideology. *American Sociological Review, 21,* 690–695.

Rost, M., Nast, R., Elger, B. S., & Shaw, D. (2021). Trust trumps comprehension, visceral factors trump all: A psychological cascade constraining informed consent to clinical trials: A qualitative study with stable patients. *Research Ethics*, *17*(1), 87-102.

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (1996, April). Empowerment in veterinary clinics: The role of trust in delegation. *In Paper presented at the symposium on trust at the 11th Annual Conference. San Diego: Society for Industrial and Organizational (SIOP).*

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An integrative model of organizational trust: Past, present, and future. *Academy of Management Review, 32*(2), 344–354.

Stone Fish, L. & Busby, D. (2005). The Delphi method. In D. Sprenkle & F. Piercy (Eds.) Research methods in family therapy (2nd edn, pp.238–253). New York: Guilford Press.

Trochim, W. M., & Donelly, J. P. (2006). Research methods base.

Turoff, M., & Linstone, H. A. (2002). The Delphi method-techniques and applications.

Ullman, D., Aladia, S., & Malle, B. F. (2021, March). Challenges and opportunities for replication science in HRI: a case study in human-robot trust. *In Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction,* (pp. 110-118).

Valenzuela, S., Park, N., & Kee, K. F. (2008, April). Lessons from Facebook: The effect of social network sites on college students' social capital. *In Paper presented at the 9th International Symposium on Online Journalism.* Texas, USA: University of Texas.

Volker Patent & Rosalind H. Searle (2019) Qualitative meta-analysis of propensity to trust measurement, *Journal of Trust Research, 9:2*, 136-163, DOI: 10.1080/21515581.2019.1675074

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion, 17,* 267-295. doi:10.1037/emo0000226

Williamson, P.R., Altman, D.G., Blazeby, J.M., Clarke, M., Devane, D., Gargon, E., & Tugwell. P. (2012). Developing core outcome sets for clinical trials: Issues to consider. Trials. 13, 132. https://doi.org/10.1186/1745-6215-13-132

Wright, E. B., Holcombe, C., & Salmon, P. (2004). Doctors' communication of trust, care, and respect in breast cancer: qualitative study. *Bmj*, *328*(7444), 864.

Wrightsman, L. (1964). Measurement of philosophies of human nature. *Psychological Reports, 14*(3), 743–751.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51*(1), 110–116.

Appendix

Table 1.

Characteristics of the experts who took part.

Characteristic	**Round 1 N (%)**	**Round 2 N (%)**
Gender		
Female	11 (39.3)	10 (43.5)
Male	17 (60.7)	13 (56.5)
Field of expertise		
Human Robot Interaction	4 (14.3)	4 (17.4)
Individual Differences	16 (57.1)	13 (56.5)
Occupational Psychology	1 (3.6)	0 (0.0)
Other[1]	7 (25.0)	6 (26.1)
Academic position		
Ph.D. Candidate	9 (32.1)	8 (34.8)
Postdoctoral Researcher	1 (3.6)	1 (4.3)
Lecturer	5 (17.9)	3 (13.0)
Senior Lecturer	4 (14.3)	3 (13.0)
Associate Professor	2 (7.1)	2 (8.7)
Professor	2 (7.1)	2 (8.7)
Other[4]	5 (17.9)	4 (17.4)

[1] For additional information regarding the expert participants, see the online repository.

Total N	**28**	**23**

Table 2.

Item relevance for Round 1.

Item	**%**	**%**	**Mean**	**Median**
	Irrelevant	**Relevant**	**(SD)**	**(range)**
I expect other people to be honest and open*.	7.14	82.14	4.18 (0.90)	4 (2-5)
I feel that other people can be relied upon. to do what they say they will do.	3.57	89.29	4.21 (0.74)	4 (2-5)
Most people are fair in their dealings with others.	10.71	75.00	3.89 (0.92)	4 (2-5)
Most people are basically honest.	0.00	92.86	4.29 (0.60)	4 (3-5)
I suspect hidden motives in others.	14.29	82.14	4.00 (1.19)	4 (1-5)
I tend to count upon other people.	7.14	67.86	3.75 (0.80)	4 (2-5)
I feel that people are generally reliable.	3.57	85.71	4.04 (0.69)	4 (2-5)
Other people cannot be relied upon.	14.29	67.86	3.75 (0.97)	4 (2-5)
Most people can be counted on to do what they say they will do.	0.00	96.43	4.36 (0.56)	4 (3-5)

I have faith in the promises or statements of other people*.	3.57	89.29	4.29 (0.76)	4 (2-5)
I believe that others have good intentions.	7.14	89.29	4.14 (0.93)	4 (1-5)
I think people generally try to back up their words with their actions.	7.14	60.71	3.68 (0.82)	4 (2-5)
Most people are honest in their dealings with others.	0.00	78.57	4.14 (0.76)	4 (3-5)
I generally give people the benefit of the doubt when I first meet them.	7.14	78.57	4.11 (0.92)	4 (2-5)
I believe that people usually keep their promises.	0.00	92.86	4.32 (0.61)	4 (3-5)
People try to take advantage of you if they get the chance.	17.86	75.00	3.86 (1.15)	4 (1-5)
People try to be fair.	3.57	50.00	3.54 (0.69)	3.5 (2-5)
People try to be helpful.	21.43	32.14	3.07 (0.94)	3 (1-5)
Other people are out to get as much as they can for themselves.	21.43	53.57	3.43 (1.17)	4 (1-5)
Other people are primarily interested in their own welfare despite what they say.	17.86	53.57	3.50 (1.07)	4 (1-5)
Other people lie to get ahead.	17.86	67.86	3.61 (1.03)	4 (1-5)

In dealing with strangers, one is better off to be cautious until they have provided evidence that they are trustworthy*.	7.14	78.57	3.93 (0.81)	4 (2-5)
I have faith in human nature.	17.86	64.29	3.61 (1.07)	4 (1-5)
I feel that other people are out to get as much as they can for themselves.	17.86	60.71	3.57 (1.07)	4 (1-5)
Most people don’t really care about what happens to the next person.	14.29	32.14	3.18 (0.67)	3 (2-4)
I trust what people say.	3.57	92.86	4.46 (0.74)	5 (2-5)
I am wary of others.	7.14	64.29	3.89 (0.96)	4 (2-5)
Most of the time, people care enough to try to be helpful, rather than just looking out for themselves*.	21.43	50.00	3.39 (1.07)	3.5 (1-5)
In general, most people keep their promises.	7.14	75.00	3.96 (0.88)	4 (2-5)
You can’t be too careful in dealing with people.	10.71	57.14	3.57 (0.96)	4 (1-5)
People are just looking out for themselves.	17.86	57.14	3.54 (0.96)	4 (2-5)
Other people who act in a friendly way towards me are disloyal behind my back.	25.00	57.14	3.50 (1.17)	4 (1-5)

Other people let you down.	28.57	46.43	3.11 (1.10)	3 (1-5)
Other people can be relied upon to do what they say they will do.	7.14	82.14	4.07 (0.86)	4 (2-5)
Other people do what they say they will do.	7.14	82.14	4.00 (0.82)	4 (2-5)
Most people tell a lie when they can benefit by doing so.	14.29	57.14	3.54 (1.00)	4 (1-5)
I generally trust other people unless they give me reason not to.	3.57	96.43	4.29 (0.81)	4 (1-5)
I usually trust people until they give me a reason not to trust them.	3.57	89.29	4.18 (0.86)	4 (1-5)
I have little faith in other people’s promises.	7.14	75.00	3.93 (0.86)	4 (2-5)
I am suspicious of other people's intentions.	10.71	82.14	4.04 (1.04)	4 (1-5)
Most people can be trusted.	0.00	100.00	4.50 (0.51)	4.5 (4-5)
Generally speaking, would you say that people can be trusted*.	14.29	71.43	3.82 (1.28)	4 (1-5)
Believe that most people would lie to get ahead*.	14.29	53.57	3.43 (1.03)	4 (1-5)
Other people live by the idea that honesty is the best policy*.	7.14	46.43	3.50 (0.92)	3 (1-5)

The typical person is sincerely concerned about the problems of others.	50.00	32.14	2.71 (1.24)	2.5 (1-5)
I generally trust other people.	0.00	96.43	4.64 (0.56)	5 (3-5)
Trusting another person is not difficult for me.	0.00	92.86	4.46 (0.64)	5 (3-5)

*Indicates items that were refined between rounds

Table 3.

Items refined based on qualitative feedback between Delphi rounds and reasons for changing items.

Original wording	**Refined wording**	**Reasons for changing item**
I expect other people to be honest and open.	I expect other people to be honest.	Original item was double-barrelled, further we are measuring trust not openness.
I have faith in the promises or statements of other people.	I have faith in the promises of other people.	Original item was double-barrelled and unnecessarily complicated.
Generally speaking, would you say that people can be trusted?	Generally speaking, people can be trusted.	Item was reworded from question format to a statement in line with other items.

Believe that most people would lie to get ahead.	I believe that most people would lie to get ahead.	Item could not stand alone, adding an ‘I’ to the item makes it stand alone and is in line with other items.
Other people live by the idea that honesty is the best policy.	Generally, people live by the idea that honesty is the best policy.	Indicates honesty is a variable rather than fixed principle of living.
Most of the time, people care enough to try to be helpful, rather than just looking out for themselves.	Most of the time, people try to be helpful.	Item was shortened and simplified in line with expert feedback.
In dealing with strangers one is better off to be cautious until they have provided evidence that they are trustworthy.	It is better to be cautious with strangers until they have shown they are trustworthy.	Item was shortened and simplified in line with expert feedback.

Table 4.

Item relevance for Round 2.

Item	% Irrelevant	% Relevant	Mean (SD)	Median (range)
I expect other people to be honest.	8.70	73.91	3.91 (0.90)	4 (2-5)
I tend to count upon other people.	26.09	52.17	3.39 (1.16)	4 (1-5)

Other people cannot be relied upon*.	21.74	65.22	3.61 (1.16)	4 (1-5)
I have faith in the promises of other people.	0.00	86.96	4.22 (0.67)	4 (3-5)
I think people generally try to back up their words with their actions.	17.39	52.17	3.35 (0.93)	4 (1-5)
People try to take advantage of you if they got the chance.	13.04	60.87	3.70 (1.11)	4 (1-5)
People try to be fair.	26.09	39.13	3.17 (0.89)	3 (2-5)
People try to be helpful.	34.78	43.48	2.96 (1.22)	3 (1-5)
Other people are out to get as much as they can for themselves.	21.74	43.48	3.26 (0.86)	3 (2-5)
Other people are primarily interested in their own welfare despite what they say.	30.43	52.17	3.26 (0.96)	4 (2-5)
Other people lie to get ahead*.	8.70	69.57	3.74 (0.81)	4 (2-5)
It is better to be cautious with strangers until they have shown they are trustworthy.	8.70	73.91	3.91 (0.90)	4 (2-5)
I have faith in human nature.	17.39	60.87	3.70 (1.06)	4 (2-5)
I feel that other people are out to get as much as they can for themselves.	26.09	52.17	3.26 (0.86)	4 (2-4)
Most people don't really care about what happens to the next person.	39.13	43.48	3.00 (1.00)	3 (1-4)
I am wary of others*.	26.09	60.87	3.57 (1.12)	4 (2-5)

Most of the time, people try to be helpful.	26.09	43.48	3.17 (1.11)	3 (1-5)
You can't be too careful in dealing with people*.	8.70	69.57	3.87 (0.92)	4 (2-5)
People are just looking out for themselves.	13.04	47.83	3.39 (0.94)	3 (1-5)
Other people who act in a friendly way towards me are disloyal behind my back*.	26.09	73.91	3.61 (10.3)	4 (2-5)
Other people let you down	30.43	39.13	3.09 (1.12)	3 (1-5)
Most people tell a lie when they can benefit by doing so*.	17.39	65.22	3.74 (1.05)	4 (2-5)
Generally speaking, people can be trusted.	0.00	100.00	4.61 (0.50)	5 (4-5)
I believe that most people would lie to get ahead.	8.70	78.26	3.78 (0.74)	4 (2-5)
Generally, people live by the idea that honesty is the best policy.	8.70	86.96	3.91 (0.73)	4 (2-5)
The typical person is sincerely concerned about the problems of others.	39.13	34.78	2.83 (1.19)	3 (1-5)

*Indicates item was retained despite not reaching consensus

List of final 33 items retained:

1. I suspect hidden motives in others.

2. I have little faith in other people’s promises.
3. I am suspicious of other people's intentions.
4. I believe that most people would lie to get ahead.
5. Other people cannot be relied upon.
6. Other people lie to get ahead.
7. I am wary of others.
8. You can't be too careful in dealing with people.
9. Other people who act in a friendly way towards me are disloyal behind my back.
10. Most people tell a lie when they can benefit by doing so.
11. It is better to be cautious with strangers until they have shown they are trustworthy.
12. I feel that other people can be relied upon to do what they say they will do.
13. Most people can be counted on to do what they say they will do.
14. Other people can be relied upon to do what they say they will do.
15. Other people do what they say they will do.
16. I have faith in the promises of other people.
17. I feel that people are generally reliable.
18. I believe that people usually keep their promises.
19. In general, most people keep their promises.
20. Most people are honest in their dealings with others.
21. Generally, people live by the idea that honesty is the best policy.
22. Most people are basically honest.
23. I expect other people to be honest.
24. Most people are fair in their dealings with others.
25. I believe that others have good intentions.

26. I generally give people the benefit of the doubt when I first meet them.
27. I generally trust other people unless they give me reason not to.
28. I usually trust people until they give me a reason not to trust them.
29. I trust what people say.
30. Most people can be trusted.
31. I generally trust other people.
32. Trusting another person is not difficult for me.
33. Generally speaking, people can be trusted.