FACULTY OF PSYCHOLOGY
AND EDUCATIONAL SCIENCES

A psychometric analysis of choice reaction time measures

Maarten De Schryver

Supervisor: Prof. Dr. Jan De Houwer
Co-supervisor: Prof. Dr. Yves Rosseel

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Doctor of psychology

Academic year 2017–2018

GHENT
UNIVERSITY

FACULTY OF PSYCHOLOGY
AND EDUCATIONAL SCIENCES

A psychometric analysis of choice reaction time measures

Maarten De Schryver

Supervisor: Prof. Dr. Jan De Houwer
Co-supervisor: Prof. Dr. Yves Rosseel

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Doctor of psychology

Academic year 2017–2018

GHENT
UNIVERSITY

Dankwoord

Net zoals mijnwerker in Potosí, zou ook de hedendaagse doctoraatstudie tot de meer stresserende beroepen ter wereld behoren. Ik mag dan ook best tevreden zijn dat ik het eindstadium van dit proefschrift bereikt heb. Dat het doctoraatstraject al bij al vlot verlopen is, heb ik vooral te danken aan allen die me de voorbije jaren gesteund hebben.

Vooreerst wil ik oprechte dank uitspreken aan mijn promotor, Jan De Houwer en copromotor, Yves Rosseel voor hun bereidheid om dit proefschrift te begeleiden, het delen van hun wetenschappelijke expertise, het nalezen, becommentariëren en soms rechttrekken van woorden en teksten.

Naast de promotor en copromotor, ben ik ook de andere leden van de doctoraats-begeleidingscommissie, met name Francis Tuerlinckx, Jonas Lang en Adriaan Spruyt, dankbaar voor hun gerichte vragen, suggesties en constructieve feedback.

Dank ook aan alle medeauteurs voor de aangename samenwerking aan de wetenschappelijke artikels waarop de verschillende hoofdstukken van dit proefschrift zijn gebaseerd: Dermot Barnes-Holmes, Aoife Cartwright, Jan De Neve, Sean Hughes, Ian Hussey, Olivier Thas en Helen Tibboel.

Veel dank aan Colin Smith voor de grondige taalrevisie; aan Jan Lammertyn voor het online krijgen van de R Shiny web apps; aan Ellen Van Glabeke om plezier in administratie te krijgen en aan Wouter Bosmans, Annick Van der Cruyssen, Linde Vandevelde en Sylvie Van Overmeeren voor alle logistieke ondersteuning.

Ook dank ik alle collega's, vrienden en familie waarop ik steeds kon rekenen. In het bijzonder Sofie voor zowat alles, Lauda en Anais om me er dagelijks aan te herinneren dat een proefschrift schrijven geen 9 to 5 maar eveneens een 5 to 9 job is.

Tenslotte wil ik ook alle medeauteurs sinds het begin van m'n academische carrière bedanken. Elk van jullie heeft vorm gegeven aan de wetenschapper die ik vandaag geworden ben.

Maarten De Schryver

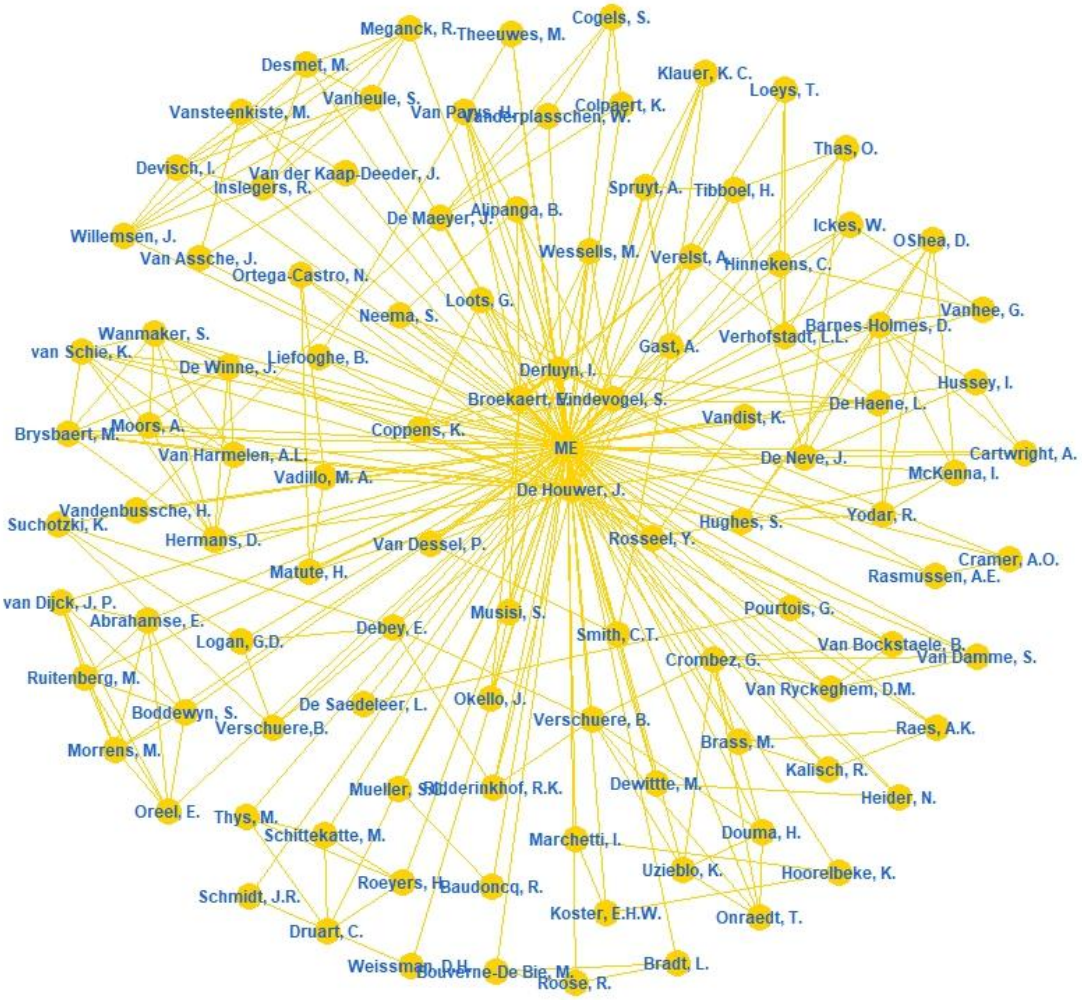

Cogels, S.
Meganck, R.
Theeuwes, M.
Desmet, M.
Klauer, K. C.
Loeys, T.
Vanheule, S.
Vansteenkiste, M.
Van Parys, H.
Colpaert, K.
Vanderplasschen, W.
Thas, O.
Devisch, I.
Van der Kaap-Deeder, J.
Inslegers, R.
Spruyt, A.
Tibboel, H.
De Maeyer, J.
Alpanga, B.
Ickes, W.
Willemsen, J.
OShea, D.
Van Assche, J.
Wessells, M.
Verelst, A.
Hinnekens, C.
Ortega-Castro, N.
Loots, G.
Vanhee, G.
Neema, S.
Barnes-Holmes, D.
Wanmaker, S.
Verhofstadt, L.L.
van Schie, K.
Gast, A.
Liefooghe, B.
De Winne, J.
Derluyn, I.
Hussey, I.
Broekaert, E.
Vindevogel, S.
De Haene, L.
Vandist, K.
Moors, A.
Brysbaert, M.
Coppens, K.
ME
Cartwright, A.
Van Harmelen, A.L.
De Neve, J.
McKenna, I.
Vadillo, M. A.
De Houwer, J.
Vandenbussche, H.
Yodar, R.
Suchotzki, K.
Hughes, S.
Hermans, D.
Van Dessel, P.
Rosseel, Y.
Cramer, A.O.
Matute, H.
Rasmussen, A.E.
van Dijck, J. P.
Pourtois, G.
Abrahamse, E.
Musisi, S.
Debey, E.
Smith, C.T.
Logan, G.D.
Van Bockstaele, B.
Van Damme, S.
Crombez, G.
Ruitenberg, M.
Van Ryckeghem, D.M.
Okello, J.
De Saedeleer, L.
Boddewyn, S.
Verschuere, B.
Raes, A.K.
Verschuere,B.
Brass, M.
Morrens, M.
Kalisch, R.
Dewittte, M.
Mueller, S.
Ridderinkhof, R.K.
Heider, N.
Oreel, E.
Thys, M.
Douma, H.
Schittekatte, M.
Marchetti, I.
Uzieblo, K.
Hoorelbeke, K.
Roeyers, H.
Baudoncq, R.
Schmidt, J.R.
Koster, E.H.W.
Onraedt, T.
Druart, C.
Weissman, D.H.
Bradt, L.
Bouverne-De Bie, M.
Roose, R.

Table of Contents

Chapter 1

General Introduction

Choice Reaction Time (CRT) measures are psychological tests that are typically derived from performance in computer-based tasks, whereby participants are instructed to react to stimuli by choosing between two or more response alternatives. Over the last two decades, their popularity has increased dramatically, most likely under the impetus of the introduction of implicit measures like, among others, the Implicit Association Test (IAT). The introduction of CRT measures raised new psychometric challenges. Despite their popularity, many statistical and psychometrical questions remain unanswered. The aim of this doctoral thesis is to further explore the psychometric properties of choice reaction time measures and to enhance these properties where possible by using insight of modern test theory and newly developed statistical models. In what follows, we start with a brief introduction of psychometrics and measurement in psychology. Next, a short (historical) overview of the use of response time in psychology is given. Subsequently, implicit measures are described and the IAT is discussed. We then reflect on potential challenges involved and present the overall objective of this doctoral thesis. We conclude this general introduction by giving an overview and a concise introduction of the remaining chapters.

To David Kinnebrook (1772 - 1802)

1.1 Psychometrics: Psychological Measurement

If she weighs the same as a duck...
she's made of wood.
And therefore?
A witch!
(From Monty Python and the Holy Grail)

In AD 932, when King Arthur and his entourage arrived in a small village, local farmers claimed to have captured a witch. After her fake nose and her hat were removed, no evidence remained to conclude that the young lady was a witch. When asked, she answered – not surprisingly – negative. Wisely, Sir Bedevere the Wise came up with some logical statements: "a witch burns, wood burns, so a witch must be made of wood, wood floats on water, ducks also float on water, so… ", and the person and a duck were put on a large scale.

The witch-scene from *Monty Python and the Holy Grail* allows us to explain and illustrate some basic concepts related to measurement in psychology. Like most *psychological constructs*, the construct "being a witch" is a *latent*, non-observable variable or trait and can only be examined via a *proxy* or *observable* data. The theory or logical statements linking the construct and the observed behavior is described in a *nomological network* (Cronbach & Meehl, 1955). This nomological network allows us to create a *measurement procedure* and a *test* based on that procedure and its result can be summarized in one single score or *measure* (the outcome of the test). In our example, the test consists of a large weighing scale and a duck, and the outcome can take on two values: "Yes" (when the person weighs the same as a duck) or "No".

Exploring the *reliability* (*the degree of measurement precision)* was rather difficult at during the early middle ages, primarily because of a high drop-out of participants (witches were burned) and because using a test re-test or alternative tests (e.g., swimming test: sink or float?; prayer test: how many mistakes?) did not work out. Also, because of the single item (the person was only measured once), internal consistency could not serve as a proxy for reliability.

Another important *psychometric* concept is that of *validity*. A test is considered *valid* if one can show that variations in the latent construct cause variations in the test scores (Borsboom, Mellenbergh and Van Heerden, 2004). A precondition of this definition is that a test can only be valid if the ontological claim (ontological claims can be objective or subjective – see Maul, 2013 for a detailed discussion) is made that the latent construct exists. A seemingly related, but different concept, is *validation*. Validation is about epistemology (i.e., the act of finding out about reality) and might – at best – provide meaning to a test (Borsboom, et al., 2004). It would take us too long to describe all types of 'validity', but some important types are *content validity* (does the test comprise all dimensions of being a witch?), *convergent validity* (do scores coherently relate to scores obtained from a witchcraft test?), *discriminant validity* (do scores not correlate with social desirability variables or Bayesian statistical skills?), *predictive validity* (can the test predict if a participant is able to turn someone into a newt?) and *construct validity* (summary of different types of validity). Note that in this doctoral thesis, validity is often also referred to by the term *utility*, to stress the difference between validity and validation. For instance, whenever the term predictive validity is used, one should realize that it refers to the results of a validation process. Moreover, utility allows us to interpret test scores also from a more pragmatic perspective.

Sir Bedevere the Wise also used another measurement procedure before putting the person and the duck on the scale. Indeed, by simply asking the question *"Is it true, are you a witch?"*, he performed a seemingly easy and simple test when asking participants to respond openly to questions: a *direct* procedure to obtain an *explicit* measure (De Houwer, 2006; De Houwer & Moors, 2010). The obtained scores can be deemed explicit because participants might carefully consider the answer. In this case, the person could be a witch – but can realize that it's opportune to answer "*No*" on this single item test. Even at that time, some people were aware about the possibility of deception and opted for other, *indirect* procedures like the duck –scale test. In contrast to direct procedures, the obtained value of an indirect procedure is subject to an interpretation of the test developer or researcher: one has to assume that the procedure will reflect the latent construct of interest.

1.2 Reaction Time Measures and the two disciplines of scientific psychology

In 1957, Lee J. Cronbach had a clear message to the American Psychological Association: the job of science is to put questions to nature – but nature will only answer if the two disciplines of scientific psychology ask their questions in a single voice. The two disciplines Cronbach was referring to are experimental psychology and correlational or differential psychology. At the time of his writing, psychologists of both streams grown apart in their interests and each discipline developed its own characteristics.

Of course, you can only grow apart if you were once united. Both disciplines started as *quantitative* sciences by shared interests in mental chronometry, that is., the empirical study of reaction time (RT) (Jensen, 2006). In the first chapter of his book "Clocking the mind: mental chronometry and individual differences", Jensen (2006) nicely describes the history of mental chronometry. A new area in psychological research was introduced by (1) the famous article of Franciscus Donders' 'On the speed of mental processes' published in 1886, (2) the experimental lab of Wilhelm Wundt at Leipzig and (3) the many chronometric studies of Wundt's student James McKeen Cattell. The ambition of the first experimental psychologists was to come up with general laws about human nature. These laws can be summarized by a simple equation R = f(S), indicating a functional relation (f) between response (R) and stimulus (S). The formula illustrates that researchers were mainly interested in effects of treatment and/or specific manipulations. As such, differences or variance between individuals were considered as nuisance or error.

Inspired by Charles Darwin's book 'On the origin of species' and by the experiments of Wundt and colleagues, Sir Francis Galton set up the first studies using RT measures for exploring individual differences. Instead of eschewing between-subject variance, Galton embraced it. As such, the above formula was extended by adding the organism (O) to it: R = f(S & O). In other words, Galton's focus was on describing how individual differences lead to different responses, given the same set of stimuli. In one of his research lines, Galton's aim was to predict occupational category based on visual and auditory RTs. By the end of 1890, he had collected data from more than 10,000 people. Unfortunately, Galton did not succeed in finding any relation between the RTs and occupational category. The consequences for Galton

were no extra publications and a research line that came to its end. Ten years later, Wissler, a PhD Student of Cattell, conducted a similar study as Galton's and collected RT data from students in Columbia University. Again, no relation was found between the RT measures and some criterion variables (in this case the students' course grades).

It is remarkable that the failures of Galton and Wissler have hampered the use of RTs in correlational psychology. Reanalysis of their data a century later, has shown that the data did not show any evidence of the absence of possible relations with the intended criteria. Rather, the data indicated a lack of evidence. First, Galton was unfortunate to have no access to later developments in statistics: analyzing his data using analysis of variances (ANOVA), invented by Fisher around 1921, revealed some *significant* results in line with Galton's hypothesis. Second, psychometric analyses, using new insights from the Classical Test Theory (starting with Spearman's correction for attenuation, in 1904), showed that the studies of Galton and Wissler both suffered from unreliable data. For instance, Galton measured only one RT (one trial) for every participant – while today researchers are recommended to measure over 60 trials (see for instance, Jensen, 2006; Miller & Ulrich, 2013). As a consequence, Galton obtained relatively stable scores when aggregating data among participants, but highly different test retest scores suggesting instability. If Cattell would have had knowledge of the concept of '*restriction of range*', he would almost certainly have given Wissler the advice to sample from a more heterogeneous population. The lack of variation in the student population of Columbia University made it almost impossible to find meaningful correlations with respect to cognitive differences among students (besides the lack of generalization, this is another drawback of running experiments only with WEIRDo's – participants from Western, Educated, Industrialized, Rich, and Democratic cultures, e.g., Jones, 2010). This brief throwback to the earlier days of scientific psychology illustrates nicely how the absence of both statistical and psychometric theory can have a direct impact on the focus of an entire discipline. These unsuccessful studies combined with the introduction of the first successful IQ tests – based on other measures than RTs – by Alfred Binet and colleagues around 1905, consigned RT measures for exploring individual differences to oblivion. It would take almost an entire century before RT measures were reintroduced for the sake of correlational research. For a detailed historical overview, I refer readers to Jensen (2006).

1.3 Implicit measures

RT measures gained renewed attention from correlational researchers by the end of the twentieth century through an increasing interest in the mediating role of cognitive processes for social psychological phenomena (Fazio, 1990). Compared to the existing tools at that time, RT measures share an interesting feature: they are based on a (seemingly) simple principle that all that should be done is measuring the time between stimulus onset and response. Also, cognitive and sensory psychologists had already extensively explored the impact of several task parameters on RTs, such as the presence of feedback, the intensity and duration of the stimuli and the clarity of instructions (Fazio, 1990), thus facilitating experimental design. Interestingly, several studies provided evidence that RT measures can be used for measuring and exploring associative network representations (Fazio, 1990). Soon thereafter, the use of RT measures was considered as an indispensable methodological tool for researchers in social psychology.

RT measures gained popularity because they seemed to be less susceptible to the problems of traditional explicit self-reports which were criticized for the fact that they strongly depend on the willingness and ability of respondents to report attributes or behavior. For instance, socially desirable or strategic responding, as well as respondents' unawareness of the construct of interest are well-known context factors that can bias measures based on self-reports (Wittenbrink & Schwarz, 2007). In response to this, implicit measures were developed. While response behavior in the context of explicit self-reports is under direct control of participants, implicit measures are assumed to capture uncontrolled and unintentional response tendencies, which are considered as a proxy of a particular attribute. The seminal work of Fazio, Jackson, Dunton and Williams (1995) using an evaluative priming task for measuring attitudes without explicit asking (see also Payne & Gawronski, 2010) and the development of the Implicit Association Test (IAT cf. infra, Greenwald, McGhee & Schwartz, 1998) may be considered as catalysts for a new class of test procedures. Today, researchers are using choice reaction time measures for investigating individual differences within many different fields, including psychopathology (e.g., Williams, Watts, MacLeod, & Mathews, 1988; Nijhof, Brass, Bardi, & Wiersema, 2016), lie research (e.g., Debey, De Schryver, Logan, Suchotzki, & Verschuere, 2014) marketing research (e.g., Richetin, Perugini, Prestwich, & O'Gorman, 2007), developmental research (e.g., Sasanguie, De Smedt, Defever, & Reynvoet, 2012).

Example: The Race-IAT The Implicit Association Test (IAT) is a computer-based procedure to measure implicit attitudes or stereotypes. The procedure, developed by Greenwald and colleagues (1998), aims to assess the relative strength of associations between two *target* categories (e.g., pictures of Black or White persons) and two *attribute* categories (e.g., pleasant or unpleasant words). In an IAT task, participants are instructed to categorize (prototypical) exemplars of these categories as fast and as accurately as possible. In a first combined phase, White and pleasant could be assigned to the same response key and black and unpleasant to a second response key. In a following phase, the two target keys are changed, while attributes remain allocated to the same key. The assumption made by the IAT developers is that the categorization task should be easier for strongly related categories compared to more weakly related categories. For instance, it might be that for some participants performance is better on the first combined task, because of a stronger association between White people and pleasant than Back people and pleasant. For participants having a stronger association between black and pleasant compared to White and pleasant, a better performance is to be expected for the second combined blocks.

Crucial for the IAT is how the score or the IAT-*effect* is calculated. First, performance is defined in terms of speed and errors (the stronger the association between two categories, the easier the task will be; hence, responding will be faster and less errors are expected). Like many choice reaction time (CRT) tasks, the effect is defined as the difference in behavior observed in the two combined phases. When the IAT was developed in 1998, the authors proposed to summarize the observed behavior using the means: for each combined block, the mean RT is calculated. The IAT-effect is then defined as the difference in mean RT. The Race-IAT participants showing positive scores are assumed to have an implicit preference for White over Black; while participants showing negative scores are assumed to have an implicit preference for Black over White. Moreover, for participants showing larger (absolute) scores, higher implicit preferences are assumed.

The IAT soon became a very popular procedure to measure all kinds of implicit preferences. Twenty years after the introduction of the IAT, the seminal paper of Greenwald et al. (1998) has been cited in over 4000 papers according to the Web of Science, and over 9000 citations according to Google Scholar. A huge amount of tests have been developed,

measuring racial bias, self-esteem, political preferences, among other things. Also, the procedure inspired many researchers to develop alternative procedures, such as the Brief-IAT (Sriram, & Greenwald, 2009), RRT (De Houwer, Heider, Spruyt, Roets, & Hughes, 2015), IRAP (Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010), EAST (De Houwer, 2003), SC-IAT (Karpinski, & Steinman, 2006), and the AMP (Payne, Cheng, Govorun, & Stewart, 2005). For more about the IAT and related procedures, I refer to Teige-Mocigemba, Klauer, and Sherman (2010).

1.4 New psychometric challenges

"Statistics requires a dynamic balance between its philosophical underpinnings and its practice to remain vital" (J. Kadane)

The introduction of CRT measures raised new psychometric challenges, for instance, related to reliability and validity of these tests (De Houwer, Teige-Mocigemba, Spruyt & Moors, 2009; LeBel & Paunonen, 2011), the parsimony of the scoring procedures used to analyze the data they generate (Blanton, Jaccard, Gonzales & Christie, 2006), and the identification of spurious factors that bias these measures (Fiedler, Messner & Bluemke, 2006). Also, the application of CRT measures as diagnostic tools in (clinical) assessment situations magnifies the need for psychometric models that take into account individual differences. Given this context, the present PhD research aims to further investigate the psychometric properties of implicit reaction time measures and develop novel approaches in testing and enhancing these properties. To this end, insights from psychometric modeling theory and new statistical methods will be integrated.

In the past, the integration of 'more advanced' psychometrics and psychological testing was not always successful. In a rather provocative (but important) paper, Borsboom (2006) describes (and illustrates) some reasons why this is the case. It would take us too far to discuss all possible reasons, but it might be useful to reflect on some of them. A first reason why researchers or test developers avoid to use more advanced methods might be related to the fact that researchers tend to rely (too much) on the Classical Test Theory. The idea of a *true score* gives researchers and test developers the illusion of a simple one-to-one mapping between a

theoretical construct and the observables. Good reliability and (the almighty) construct validity indices incorrectly give the impression that one can simply interpret results obtained from a test as a proxy of the theoretical variable of interest. Another reason given by Borsboom (2006) is more substantial: psychologists seem to prefer pragmatism (or convention) above more fine-grained psychometric theories. Other reasons mentioned by Borsboom (2006) are more pragmatic in nature. The availability of (statistical) software (e.g., SPSS allows the user to easily calculate Cronbach's alpha, but not more advanced reliability indices), a lack of mathematical training in psychology ("Every trained economist understands basic calculus, for instance, while trained psychologists often do not know what calculus is in the first place.", p. 432), and the apparent normative character of psychometric theory, are instances of pragmatic reasons discussed by Borsboom (2006).

Therefore, it is important to keep in mind the somewhat troubled relation between psychology and psychometric theory and the possible reasons why this is the case if we want to reach our set objectives. Investigating psychometric properties and exploring new statistical methods can be interesting for the psychometrician, but one of the goals should be to offer (applied) researchers information that is comprehensive without being too technical. Also, researchers should be offered tutorials or software to facilitate the integration of psychometric theory, advanced statistical modeling and psychological theory. Psychometric studies are only meaningful when a) test users are convinced that there is a merit in changing their current practices and b) they know how to implement new methods.

1.5 Research Objectives and Overview of the chapters

Although the principle of measuring the time between stimulus onset and response might appear simple, using latency or RT data in a measurement context is far more complex. First, RT data contain a rather low signal-to-noise ratio and vary substantially between tasks, experimental conditions and participants. Individual differences are observed in speed-accuracy trade-off and general response speed (GRS). Differences in GRS seem to be an important confounder in exploring individual differences between conditions (Fazio, 1990; Faust, Balota, Spieler, and Ferraro, 1999; Greenwald, et al., 2003). For instance, when an effect is defined as a difference in mean RT between congruent and incongruent conditions (e.g.,

which was initially the case for the IAT), a strong positive correlation is observed between GRS and the effect itself. Moreover, calculating or estimating effects would be much easier if RT distributions are normal. However, it is well known that these distributions can be heavily skewed and outliers can be expected (Fazio, 1990; Van Zandt, 2002; Wagenmakers & Brown, 2007; Balota and Yap, 2011). This might result in biased estimates of the mean and variance, hence influencing the outcome of the measure itself.

Recently, Thas, De Neve, Clement, and Ottoy (2012) introduced the Probabilistic Index Models (PIM's). This is a new class of semiparametric regression models. PIM's, which are robust to outliers, might function as a promising new framework to define differences in reaction time performance in terms of the probability that a response latency increases if the context changes. In **Chapter 2**, a general introduction to these models is provided because, to date, only (mathematically) technical papers exist and the applications of PIMs are mainly situated in the health and biological sciences. In this chapter, we provide an introduction to PIMs by 1) discussing key features of the model, 2) motivating why we think PIMs could be useful for behavioral sciences, 3) demonstrating PIMs on a case study and 4) illustrating how PIMs can be used in practice using the R package `pim`.

Next, alternative scoring algorithms based on PIMs are proposed for the IAT (**Chapter 3**) and for the Implicit Relational Assessment Procedure (IRAP; **Chapter 4**). Although both procedures stem from different research traditions, they both share – in principle – a similar effect size measure, which can be called the D-effect sizes measure. The D-effect size measure is part of the D-scoring algorithms, and was proposed by Greenwald and colleagues (2003) for scoring IAT data. The D-scoring algorithms have so far been considered as the 'best' way to calculate individual IAT scores, both for IAT data collected on the internet (Greenwald, et al., 2003; Richetin, et al., 2015) and for data collected in a laboratory setting (Glashouwer, et al., 2013; Richetin, et al., 2015). Based on existing criteria and some new criteria reflecting statistical properties, the newly proposed scoring algorithms and existing scoring algorithms are evaluated and compared.

The second part of this dissertation focuses on reliability. Although reliability is generally considered as one of the most important psychometric properties of a psychological test, a

general framework for estimating or approximating and interpreting reliability is lacking. The low signal-to-noise ratio of test procedures based on RTs and the substantial presence of *error* variance have been major concerns for using RT-based implicit measures [1]and have been reflected in low to moderate *reliability* scores (e.g., Bosson, Swann & Pennebaker, 2000; Kawakami & Dovidio, 2001; LeBel & Paunonen, 2011). Based on these observations, low reliability might become the Achilles' heel of implicit measures. Moreover, while many researchers might be familiar with the reliability concept, research has shown that due to ignorance of measurement theory and practice, only a quarter of doctoral students were able to correctly apply methods of reliability (e.g. Aiken et al., 1990; Graham, 2006).

In **Chapter 5**, we provide a tutorial that has two goals. (1) It provides a quick primer for those interested in the concept of reliability and its relationship to RT-based implicit measures. In doing so, we start by clarifying some conceptual issues, such as the difference between a *test* and a *procedure*. We also argue that reliability can only be approximated via *consistency*, *equivalence*, and *stability*, whereby some elements in the context are manipulated. (2) It introduces a general framework, using a Latent Variable Model – or more specific – a Confirmatory Factor Analysis (CFA) framework that can be used by novel and seasoned researchers alike to estimate and interpret the reliability of their implicit measures. This framework goes beyond a strict Classical Test Theory by offering more flexibility by (explicitly) representing attributes as latent variables.

Finally, in **Chapter 6**, we state that arguments and recommendations used by LeBel and Paunonen (2011) regarding the role of reliability are problematic and that they might undermine the interpretation and evaluation of empirical findings as well as the development of new procedures. In their original paper, LeBel and Paunonen (2011) drew attention to the implicit measurement revolution that has unfolded within social psychology. The authors argued that these measures suffer from unacceptably low levels of reliability and suggested that lower levels of reliability are directly associated with decreasing probabilities of replicating an effect. As a consequence, they state that researchers should strive to improve implicit measures that fall prey to unacceptable levels of reliability or utilize measures known

[1] Note that not all CRT-measures are implicit measures and not all implicit measures are based on CRT. In this dissertation the main focus is on implicit measures based on CRT.

to have acceptable psychometric properties. They also conclude that researchers should report reliability estimates separately for each experimental condition. We argue that there is in fact no direct relation between reliability and power and/or replicability. Low reliability is not always a problem and might actually reflect tight experimental control. We address these various concerns in our commentary and offer the reader several recommendations that should be taken into account when examining the relationship between reliability, replicability and power.

References

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., Roediger III., H. L., Scarr, S., ... & Sherman, S. J. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist*, *45*(6), 721.

Balota, D. A., & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry: The power of response time distributional analyses. *Current Directions in Psychological Science*, *20*(3), 160-166.

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record*, *60*(3), 527.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, *42*(2), 192-212.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425-440.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, 111(4), 1061-1071

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited?. *Journal of personality and social psychology*, *79*(4), 631-643.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American psychologist*, *12*(11), 671-684.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281-302.

De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental psychology*, *50*(2), 77-85.

De Houwer, J. (2006). *What are implicit measures and why are we using them.* In R. W. Wiers and A. W. Stacy (Eds.), The handbook of implicit cognition and addiction (pp. 11-28). Thousand Oaks, CA: Sage Publishers.

De Houwer, J., Heider, N., Spruyt, A., Roets, A., & Hughes, S. (2015). The relational responding task: toward a new implicit measure of beliefs. Frontiers in psychology, 6.

De Houwer, J., & Moors, A. (2010). Implicit measures: Similarities and differences. In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications*. New York, NY: Guilford Press

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological bulletin*, *135*(3), 347-368.

Debey, E., De Schryver, M., Logan, G. D., Suchotzki, K., & Verschuere, B. (2015). From junior to senior Pinocchio: A cross-sectional lifespan investigation of deception. *Acta psychologica*, *160*, 58-68.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: implications for group differences in response latency. *Psychological bulletin*, *125*(6), 777.

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. *Research methods in personality and social psychology*, *11*, 74-97.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of personality and social psychology*, *69*(6), 1013.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology*, *17*(1), 74-147.

Glashouwer, K. A., Smulders, F. T., de Jong, P. J., Roefs, A., & Wiers, R. W. (2013). Measuring automatic associations: Validation of algorithms for the Implicit Association Test (IAT) in a laboratory setting. *Journal of Behavior Therapy and Experimental Psychiatry*, *44*(1), 105-113

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability what they are and how to use them. *Educational and Psychological Measurement*, *66*(6), 930-944.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.

Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences*. Elsevier.

Jones, D. (2010). A WEIRD view of human nature skews psychologists' studies. *Science*, *328*(5986), 1627-1627.

Karpinski, A., & Steinman, R. B. (2006). The Single Category Implicit Association Test as a measure of implicit social cognition. *Journal of Personality and Social Psychology*, *91*(1), 16-32.

Kawakami, K., & Dovidio, J. F. (2001). The reliability of implicit stereotyping. *Personality and Social Psychology Bulletin*, *27*(2), 212-225.

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin*, *37*(4), 570-583.

Lord, F. M., & Novick, M. R. (1967). *Statistical theories of mental test scores*. IAP.

Maul, A. (2013). On the ontology of psychological attributes. *Theory & Psychology*, *23*(6), 752-769.

Miller, J., & Ulrich, R. (2013). Mental chronometry and individual differences: Modeling reliabilities and correlations of reaction time means and effect sizes. *Psychonomic bulletin & review*, *20*(5), 819-858.

Nijhof, A. D., Brass, M., Bardi, L., & Wiersema, J.R. (2016). Measering Mentalzing Ability: A within-subject comparison between an explicit and implicit version of a ball detection task. Plos One, 11 (10), e0164373. Htpp://doi.org/10.1371/journal.pone.0164373

Payne, B., Cheng, C., Govorun, O., & Stewart, B. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89, 277–293

Payne, B. K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going. *Handbook of implicit social cognition: Measurement, theory, and applications*, *1*, 1-15.

Richetin, J., Perugini, M., Prestwich, A., & O'Gorman, R. (2007). The IAT as a predictor of food choice: The case of fruits versus snacks. *International Journal of Psychology*, *42*(3), 166-173.

Sasanguie, D., De Smedt, B., Defever, E., & Reynvoet, B. (2012). Association between basic numerical abilities and mathematics achievement. *British Journal of Developmental Psychology*, *30*(2), 344-357.

Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental psychology*, *56*(4), 283-294.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta psychologica*, *30*, 276-315.

Teige-Mocigemba, S., Klauer, K. C., & Sherman, J. W. (2010). A practical guide to the Implicit Association Test and related tasks. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 117-139). New York: Guilford Press

Thas, O., De Neve, J., Clement, L., and Ottoy, J.P. (2012). Probabilistic index models (with Discussion). *Journal of the Royal Statistical Society - Series B*, 74:623–671.

Wagenmakers, E. J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological review*, 114(3), 830.

Van Zandt, T.(2002). Analysis of response time distributions. In J. T. Wixted (Vol. Ed.) & H. Pashler (Series Ed.) *Stevens' Handbook of Experimental Psychology (3rd Edition), Volume 4: Methodology in Experimental Psychology*(pp. 461-516). New York: Wiley Press.

Williams, J. M. G., Watts, F. N., MacLeod, C., & Mathews, A. (1988). *Cognitive psychology and emotional disorders*. John Wiley & Sons.

Wittenbrink, B., & Schwarz, N. (Eds.). (2007). *Implicit measures of attitudes*. Guilford Press.

Kadane, 1976, p. 735 as cited in Serlin (2002). "statistics requires a dynamic balance between its philosophical underpinnings and its practice to remain vital". Same holds true for psychometrics! (Serlin, R. C. (2002). Constructive criticism. *Journal of Modern Applied Statistical Methods*, (2), 31.)

Chapter 2

An introduction to probabilistic index models: regression models for the effect size P (Y1 < Y2)

Maarten De Schryver and Jan De Neve

Abstract

The probabilistic index (PI), also known as the probability of superiority or the common language effect size, refers to the probability that the outcome of a randomly selected subject exceeds the outcome of another randomly selected subject, conditional on the covariate values of both subjects. This summary measure has a long history, especially for the two-sample design where the covariate value typically refers to one of two treatments. Despite some of the attractive features of the PI, it is often not used beyond the two-sample design. One reason is the lack of a flexible regression framework that embeds the PI and allows the user to estimate it for more complicated designs. However, Thas, De Neve, Clement and Ottoy (2012) recently developed such a regression framework, named Probabilistic Index Models (PIMs). In this tutorial we provide an introduction to PIMs where we discuss several theoretical properties, motivate why we think PIMs could be useful for behavioral the sciences, and illustrate how it can be used in practice using the R package `pim`.

2.1 Introduction

Probabilistic index models (PIMs) are a class of semiparametric regression models in which the probabilistic index (PI) is modeled as a function of covariates. The PI refers to the probability that the outcome of a randomly selected subject exceeds the outcome of another randomly selected subject, conditional on the covariate values of both subjects. Let Y denote the univariate outcome and $\boldsymbol{X}$ the p-dimensional vector of covariates. If $(Y_i, \boldsymbol{X}_i^T)$ denotes the observation of subject i and $(Y_j, \boldsymbol{X}_j^T)$ that of subject j, then the PI is given by $\mathrm{P}\left(Y_i < Y_j \mid \boldsymbol{X}_i, \boldsymbol{X}_j\right)$. As an illustration, let Y denote the Beck Depression Inventory (BDI) II depression score (range 0 – 63). Lower scores indicate less severe depression. Patients were randomized to one of two treatments: an innovative therapy (dummy coded as $X = 0$) or a conventional therapy ($X = 1$). The PI, defined here as $\mathrm{P}\left(Y_i < Y_j \mid X_i = 0, X_j = 1\right)$, gives the probability that a randomly selected patient of the innovative therapy group will have a lower BDI score than a randomly selected patient of the conventional therapy group. This probability can then be used as a summary measure to quantify the treatment effect.

The PI is the effect measure associated with the Wilcoxon–Mann–Whitney test, also known as the Mann–Whitney test or the Wilcoxon-rank-sum test (Wilcoxon, 1945; Mann and Whitney, 1947; Kruskal, 1952). Bamber (1975) discusses how the PI can be used as a measure of the size of the difference between two populations and as a measure of discrimination accuracy. Cliff (1993) argues that the PI is an appropriate effect measure in behavioral research and Acion Peterson, Temple and Arndt (2006) state that the PI is a simple, clinically-relevant and robust index. Ruscio (2008) illustrates that, unlike Cohen's d or the point-biserial correlation, the PI estimator is robust to base rates. The PI might especially be relevant in psychology, since the PI is unaffected by monotone transformations, making it a relevant effect measure when the observed outcome is monotonically related to the underlying latent variable (Grissom & Kim, 2001). Evidently, the PI as an effect size also has its shortcomings and has been criticized; see e.g. Senn (1997, 2006, 2011). For more reading on the PI as an effect measure and how to conduct inference in the two-sample case, we refer to, among others, Grissom (1994); Laine and Davidoff (1996); Brunner and Munzel (2000); Hauck, Hyslop and Anderson (2000); Vargha and Delaney (2000); Kotz and Pensky (2003); Newcombe (2006); D'Agostino, Campbell and Greenhouse (2006); Zhou (2008); Tian (2008); Ruscio and Mullen

(2012); Ruscio and Gera (2013); Thas, De Neve, Clement and Ottoy (2012); Kieser, Friede and Gondan (2013); De Neve and Thas (2015) and the references therein.

For the two-sample problem, $\mathrm{P}\left(Y_i < Y_j \mid X_i = 0,\ X_j = 1\right)$ has been given many names: the measure of stochastic superiority, the probability of superiority, the common language effect size, the dominance statistic, the nonparametric treatment effect, the relative treatment effect, the individual exceedance probability, the reliability or the probabilistic index, among others. We choose the latter, a term coined by Acion et al. (2006), while acknowledging that this name is not optimal (this also holds true for the other names), since the probability $\mathrm{P}\left(Y_i < Y_j \mid X_i = 0,\ X_j = 1\right)$ can be considered as *a* PI and not necessarily *the* PI.

The vast majority of articles on the PI focus on the two-sample design since the PI can then be easily estimated. Some authors, however, have extended the estimation to more complicated designs. Tian (2008) developed a parametric regression model for the PI assuming a normal linear regression model. Brumback, Pepe and Alonzo (2006) developed a semiparametric model by using methods for receiver operating characteristic (ROC) curve regression analysis to accommodate the Wilcoxon–Mann–Whitney test for covariate adjustment (Dodd & Pepe, 2003). Their methodology is, however, still restricted to two-sample designs and does not allow quantification of the effect of a continuous covariate on the outcome in terms of the PI. Thas et al. (2012) introduced a class of regression models, named Probabilistic Index Models (PIMs), where they model the PI directly as a function of the covariates and this in a semiparametric fashion. This methodology allows estimating the PI for a variety of designs, including designs with multiple and/or continuous covariates. In De Neve and Thas (2015) it is shown that many well-known nonparametric rank tests (e.g. Wilcoxon–Mann–Whitney, Kruskal–Wallis, Friedman rank tests) can be embedded in the PIM-framework in a similar fashion as how *t*- and *F*-tests can be embedded in a linear regression model. The PIM formulation further allows construction of confidence intervals for the PI in addition to hypothesis testing. This illustrates that a PIM can be seen as the natural extension of the Wilcoxon–Mann–Whitney test to a regression context, without being limited to two-sample designs.

Since the introduction of PIMs in Thas et al. (2012), several follow-up articles have been written (De Neve, Thas & Ottoy, 2013a; De Neve, Thas & Ottoy, Clement, 2013b; De Neve, Meys, Ottoy, Clement & Thas, 2014; De Neve & Thas, 2015; Vermeulen, Thas, Vansteelandt, 2015; Amorim, Thas, Vermeulen, Vansteelandt & De Neve, 2017). These articles are, however, focused on applications from the health and biological sciences and cannot be considered as tutorials. In this article we provide an introduction to PIMs with specific focus on the behavioral sciences. We discuss both theoretical and practical results and illustrate how the R package **pim** (Meys, De Neve, Sabbe & Amorim, 2017; R Core Team, 2017) can be used. We believe PIMs can provide additional insight in psychological processes under study. Therefore, we hope that this tutorial can stimulate the community to apply these models in practice.

The paper is organized as follows. In Section 2.2, we discuss some of the most important aspects of the PI in the two-sample design. Some of these results have been discussed in the literature, nevertheless we repeat them here since they are relevant for understanding PIMs. In addition to well-studied properties, we also discuss some important features that have not been given a lot of attention. In Section 2.3, we introduce PIMs by considering a univariate continuous covariate before extending it to the multi-variable setting. In Section 2.4, we illustrate the relationship between PIMs and several other models, including the normal linear regression model and the Cox proportional hazards model. In Section 2.5, we discuss goodness-of-fit assessment, and in Section 2.6 we illustrate the method on a case study and provide R code. Section 2.7 contains the conclusion.

2.2 The probabilistic index for the two-sample design

2.2.1 Effect measure beyond the mean

Normal distributions

The PI is an effect measure that captures effects beyond the mean and it is to some extent related to a standardized mean difference. This is most easily seen by considering normal distributions. We consider the randomized trial of Section 1 with Y, the BDI score, and X, the treatment. By Y_{IT} (Y_{CT}) we denote the outcome of the innovative (conventional) therapy so that

we can write the PI compactly as P ($Y_{CT} < Y_{IT}$). We further assume that both outcomes are independent and normally distributed: $Y_{IT} \sim N(\mu_{IT}, \sigma_{IT}^2)$ and $Y_{CT} \sim N(\mu_{CT}, \sigma_{CT}^2)$. To summarize the association between *Y* and *X* we can look at the mean difference $\mu_{CT} - \mu_{IT}$ or its standardized version

$$\delta = \frac{\mu_{CT} - \mu_{IT}}{\sqrt{\sigma_{CT}^2 + \sigma_{IT}^2}}.$$

It follows that the difference $Y_{IT} - Y_{CT} \sim N(\mu_{IT} - \mu_{CT}, \sigma_{IT}^2 + \sigma_{CT}^2)$. It is now straightforward to derive the PI

$$\text{P}(Y_{IT} < Y_{CT}) = \text{P}(Y_{IT} - Y_{CT} < 0) = \Phi\left(\frac{\mu_{CT} - \mu_{IT}}{\sqrt{\sigma_{CT}^2 + \sigma_{IT}^2}}\right) = \Phi(\delta)\text{ , (1)}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. We consider an example to illustrate some properties of these three effect measures.

Figure 1 (top panels) provides artificial data for two groups of patients with depression: patients that receive antidepressants (left panel) and patients that do not receive antidepressants (right panel). We first consider the left panel. The mean difference is 5 and $\delta = 0.75$. It is clear that the treatment affects both the location and the scale of the distribution. The standardized mean difference can therefore be considered as an appropriate effect measure for this setting. A disadvantage, however, is that it is more difficult to interpret than the mean difference: it states that the mean BDI score in the innovative treatment group will be 0.75 standard deviations lower as compared to the mean BDI score in the conventional treatment group. The standard deviation refers to the deviation of the differences between patients of both groups $Y_{CT} - Y_{IT}$.

The PI combines the desirable properties of the mean difference and the standardized mean difference: it accounts for the change in variability, while retaining a relevant interpretation. It holds that P ($Y_{IT} < Y_{CT}$) = $\Phi(\delta) = 77\%$. Hence there is a 77% chance that a patient of the innovative treatment group will have a lower BDI score as compared to a patient of the conventional treatment group. The importance of standardization becomes even more apparent when looking to the right panel of Figure 1. It is visually clear that the effect of the therapy is different as compared to the left panel. The mean difference, however, does not pick this up: it remains 5 and the standardized difference is $\delta = 0.37$. The PI P ($Y_{IT} < Y_{CT}$) =

65% does pick up this change and shows a decreased effect due to an increase in variability. Similar as for the left panel, the PI gives an effect measure on an interpretable scale, whereas δ can be hard to understand (e.g. Ruscio, 2008).

The interpretation of the PI can be considered an attractive property. However, it also has limitations since it is not able to distinguish between densities that are completely separated. The bottom panels of Figure 1 illustrate this. The mean difference for the patients receiving antidepressant drugs is 10, while for patients not receiving antidepressants this is 20. For the standardized differences, the effects are $\delta = 3.5$ and $\delta = 7.1$. Both the mean difference and the standardized difference indicate a larger treatment effect for the no antidepressant group. The PI, on the other hand, is not capable of quantifying this difference because it is approximately 1 for both groups, which is reflected by the non-overlap of the two densities.

From these two examples it should be clear that the mean difference, the standardized mean difference and the PI are three different measures to quantify treatment effects, each with their own strengths and weaknesses.

Skewed distributions

It is important to realize that the simple relation between δ and the PI as defined in (1) does certainly not always hold. To illustrate this, we consider the skewed log-normal (LN) distribution $Y_{IT} \sim LN(\mu_{IT}, \sigma_{IT}^2)$ and $Y_{CT} \sim LN(\mu_{CT}, \sigma_{CT}^2)$. It follows that

$$\mathrm{E}(\mathrm{Y}_{CT}) - \mathrm{E}(\mathrm{Y}_{IT}) = \exp\left(\mu_{CT} + \frac{\sigma_{CT}^2}{2}\right) - \exp\left(\mu_{IT} + \frac{\sigma_{IT}^2}{2}\right)$$

and

$$\delta = \frac{\exp\left(\mu_{CT}+\frac{\sigma_{CT}^2}{2}\right)-\exp\left(\mu_{IT}+\frac{\sigma_{IT}^2}{2}\right)}{\sqrt{\exp\left(\sigma_{CT}^2-1\right)\exp\left(2\mu_{CT}+\sigma_{CT}^2\right)+\exp\left(\sigma_{IT}^2-1\right)\exp\left(2\mu_{IT}+\sigma_{IT}^2\right)}}.$$

The log-normal distribution arises from exponentiating normal variables. Since the PI is not affected by such a transformation, see Section 2.4 for a more detailed discussion, it follows that

$$\mathrm{P}\left(Y_{IT} < Y_{CT}\right) = \Phi\left(\frac{\mu_{CT}-\mu_{IT}}{\sqrt{\sigma_{CT}^2+\sigma_{IT}^2}}\right).$$

This demonstrates that the PI is an effect measure on its own and is definitely not just a (simple) transformation of δ to an interpretable scale.

2.2.2 Interpretation

The interpretability of the PI can be considered as an attractive feature (Acion et al., 2006). There is, however, also a downside: it is an effect measure that can be easily misunderstood (Senn, 2006). In the randomized clinical trial were patients are either assigned to the innovative ($X = 0$) or the conventional therapy ($X = 1$), the PI gives the probability that a randomly selected patient of the innovative treatment group will have a lower BDI score as compared to a second patient that is randomly selected from the conventional therapy group. The PI does not give the probability that a single randomly selected patient will have a lower BDI score when given the innovative treatment as compared to the conventional treatment.

This is most easily seen by considering normal distributions. We assume as before that $Y_{IT} \sim N(\mu_{IT}, \sigma_{IT}^2)$ and $Y_{CT} \sim N(\mu_{CT}, \sigma_{CT}^2)$ so that $Y_{IT} - Y_{CT} \sim N(\mu_{IT} - \mu_{CT}, \sigma^{*2})$, where $\sigma^{*2} := \mathrm{Var}(Y_{IT} - Y_{CT})$. We now deliberately do not yet work out this term. It follows that

$$\mathrm{P}(Y_{IT} < Y_{CT}) = \Phi\left(\frac{\mu_{CT} - \mu_{IT}}{\sigma^*}\right).$$

In an independent-samples randomized trail where patients receive only one of the two treatments, Y_{CT} and Y_{IT} refer to outcomes of two different patients. Since samples are independent, it follows that $\sigma^{*2} = \sigma_{CT}^2 + \sigma_{IT}^2$ and the PI becomes

$$\mathrm{PI} = \Phi\left(\frac{\mu_{CT} - \mu_{IT}}{\sqrt{\sigma_{CT}^2 + \sigma_{IT}^2}}\right).$$

If, on the other hand, a paired design would have been considered where each patient gets both treatments, then Y_{CT} and Y_{IT} refer to the outcomes of the same patient. These outcomes are not independent so that $\sigma^{*2} = \sigma_{CT}^2 + \sigma_{IT}^2 - 2\mathrm{Cov}(Y_{CT}, Y_{IT})$. The PI then becomes

$$\mathrm{PI} = \Phi\left(\frac{\mu_{CT} - \mu_{IT}}{\sqrt{\sigma_{CT}^2 + \sigma_{IT}^2 - 2\mathrm{Cov}(Y_{CT}, Y_{IT})}}\right).$$

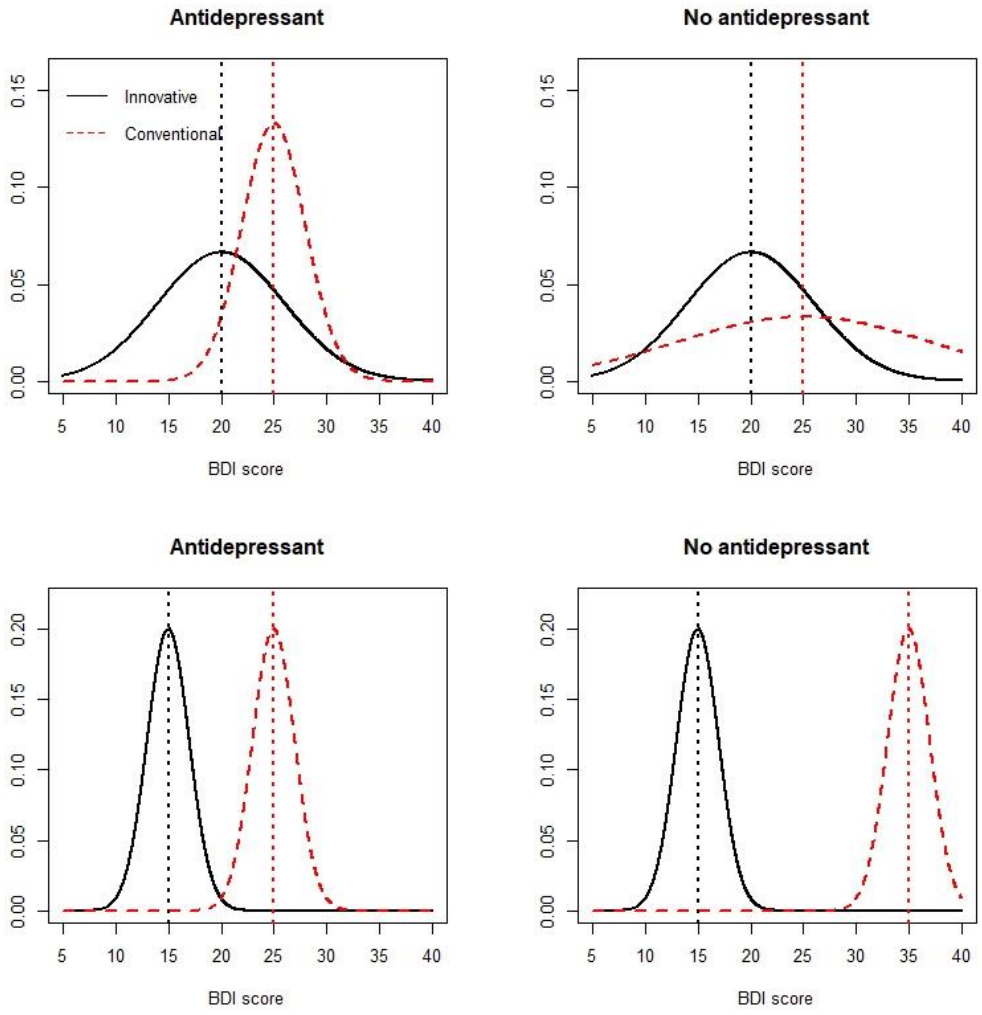


Figure 1. Artificial data where the innovative treatment outperforms the conventional treatment in terms of BDI scores for patients that receive antidepressants (left panels) or that do not receive antidepressants (right panels). Top panels: for the left panel, the standard deviation is smaller for the BDI scores of the conventional therapy, while the opposite holds for the panel on the right. The PI P $(Y_{IT} < Y_{CT})$ = 77% while for the right panel this decreases to 65%. For both panels, the mean difference is 5 (the vertical lines represent the mean of each population). The standardized mean difference equals 0.75 for the left panel and 0.37 for the right panel. Bottom: for both panels the densities show minimal overlap. The mean difference and its standardized version is smaller for the left panel, while the PI is approximately 1 for both panels.

It is clear that both PI's are not the same and it demonstrates that the interpretation of the PI depends on the design of the study: independent- or paired-samples, or more generally independent versus clustered data. In this manuscript, we focus on estimating the PI based on

identically and independently distributed (i.i.d.) data $(Y_i, \boldsymbol{X}_i^T)$ and hence we are always comparing two different subjects.

2.2.3 Gaining power by exploiting order

The most well-known estimator of the PI is the one that forms the basis of the non-parametric (or rather distribution-free) Wilcoxon–Mann–Whitney (WMW) test. If we have n_{IT} subjects receiving innovative therapy and n_{CT} subjects receiving conventional therapy, the PI can be unbiasedly estimated via the Mann–Whitney statistic

$$MW = \frac{1}{n_{IT} n_{CT}} \sum_{i=1}^{n_{IT}} \sum_{j=1}^{n_{CT}} \mathrm{I}(Y_{IT,i} < Y_{CT,j}), \quad (2)$$

where $\mathrm{I}(\cdot)$ denotes the indicator function for which $\mathrm{I}\left(Y_{IT,i} < Y_{CT,j}\right) = 1$ if $Y_{IT,i} < Y_{CT,j}$ and $\mathrm{I}\left(Y_{IT,i} < Y_{CT,j}\right) = 0$ otherwise (ties are addressed in Section 3.3). Lehmann (1951) showed that MW is the uniform minimum variance unbiased estimator of $\mathrm{P}(Y_{IT} < Y_{CT})$ within a large class of continuous distributions (Thas, 2010). The WMW test statistic is then given by

$$\sqrt{\frac{12 n_{IT} n_{CT}}{n_{IT} + n_{CT} + 1}} (WM - 0.5).$$

This expression makes it clear that the PI is the effect size associated with the WMW test. At first sight, it might seem that we are losing information in (2) by only considering the relative ordering of the outcomes and by ignoring the magnitude of the differences (as illustrated in the lower panels of Figure 1). However, it is well known that performing a hypothesis test based on statistic (2) can lead to a substantial gain in power as compared to a test based on the difference in sample means. Table 1 shows the asymptotic relative efficiency (ARE) of the t-test relative to the WMW test when both groups only differ in terms of their location. Under normality, the t-test is superior (i.e. ARE is less than one) since its parametric assumptions are fulfilled. Notice, however, that the superiority is rather modest. Roughly speaking, using the t-test one only requires 95% of the observations as compared to using the WMW test to achieve the same power.

There are, however, many distributions for which the WMW test is (substantially) superior. If observations are coming from a skewed distribution such as the exponential, the t-test needs 3 times as many observations as the WMW test to achieve the same power

(asymptotically speaking). It is worth mentioning that the superior performance of the t-test is bounded: it cannot perform better than the setting where the density is $f(y) = \max(1 - y^2, 0)$, while the superiority of the WMW test is unbounded when data are coming from the heavy-tailed Cauchy distribution. We refer to Lehmann (2004) for a detailed discussion on the ARE.

To illustrate that the WMW test can also be superior in small samples (thus not relying on asymptotic results), Figure 2 gives the power (approximated based on 10000 Monte-Carlo simulations) for balanced two-sample designs (with 20 or 40 observations per group), where the mean difference between both groups equals a half standard deviation. A permutation null distribution is used for both tests so that deviations from normality do not invalidate the statistical inference. We can see that even for small samples, the WMW test can outperform the t-test in terms of power.

Table 1. Asymptotic relative efficiency (ARE) of the *t*-test versus the WMW test for different distributions. A value larger than 1 implies that the WMW is (asymptotically) more efficient.

Distribution:	max(1-y^2,0)	Normal	Uniform	Logistic	t_3	Laplace	t_5	Exp	Cauchy
ARE:	0.86	0.95	1	1.1	1.24	1.5	1.9	3	∞

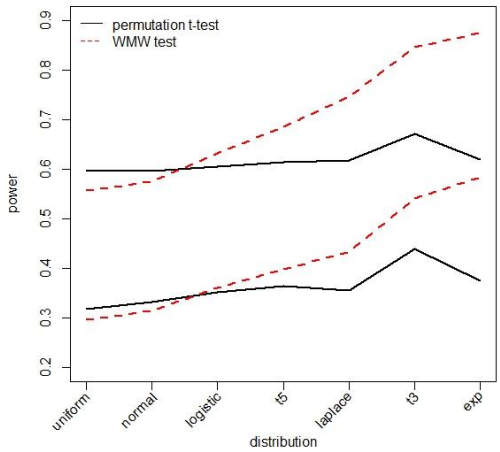


Figure 2. Power for a balanced two-sample design with 20 (lower lines) or 40 (upper lines) observations per group and for several choices of distributions. Both groups have the same distribution except for half standard deviation difference in location.

2.2.4 Invariance under monotone transformations

Let $h(\cdot)$ denote a strictly increasing monotone function (e.g. the exponential), then $\mathrm{P}\,(h(Y_i) < h(Y_j) | X_i = 0, X_j = 1) = \mathrm{P}\,(Y_i < Y_j | X_i = 0, X_j = 1)$, i.e. the PI is invariant under monotone transformations. This property could make the PI an attractive effect measure in psychology, where the observed data are typically used to study an underlying latent psychological construct. Because this construct can never be observed, it is practically impossible to specify the exact relationship between the observed data and the construct (Grissom & Kim, 2005). If you are willing to assume that this relationship is monotone (without the need to specify the exact form of the function), the PI allows the researcher to formulate conclusions about this construct based on the observed data. This is especially relevant for non-linear monotone functions, because they can make statistical interactions appear or disappear when interactions are defined in terms of differences in means (Loftus, 1978; Wagenmakers, Krypotos, Criss & Iverson, 2012; Garcia-Marques, Garcia-Marques & Brauer, 2014).

We consider an example to illustrate this. Let Y denote the BDI score and let θ denote the latent variable of interest. The middle panel of Figure 3 shows the relationship between Y and θ. The curvelinearity implies that patients are not reporting extreme feelings, e.g. for social desirability reasons (Garcia-Marques et al., 2014). Even if a psychologist is unwilling to define a univariate underlying psychological construct of interest, θ can be considered as the BDI score when patients would express extreme feelings. We assume that the monotone transformation is psychometrically invariant, i.e. it is not affected by the treatments (Davison & Sharma, 1990).

When we quantify effects via the mean difference, there is no interaction for the latent construct: the mean difference is 5 for both groups (antidepressant and no antidepressant). The PI equals 88% in both groups, also indicating that the effect size does not depend on whether patients receive antidepressants. When applied to the BDI scores, the PI's are unaffected since they are monotone invariant. The mean differences, however, have changed. For the antidepressant group, the mean difference decreased to 4.6, while for the group not receiving antidepressants the effect decreased to 2.7. This decrease is caused by the curvilinearity of the

relationship between the latent construct and the BDI score. The decrease is larger for patients receiving no antidepressants since they report, in general, higher scores and hence are more affected by the decrease in slope of the monotone function. Based on the observed BDI scores we would conclude that antidepressants act as a moderator, while this is not true for the latent construct. This interaction is an artifact of the nonlinear relationship between the latent construct and the observed outcome. When we quantify the effect in terms of the PI instead of in terms of the difference in means, we would conclude that antidepressants do not act as a moderator. This conclusion holds true for both the observed score and the latent variable.

It is important to mention that these properties only hold under the restrictive setting where we assume a one-to-one mapping between the latent construct and the BDI score. In practice, it is more realistic to allow for measurement error and then the PI does not allow one to directly translate the conclusions from the BDI scores to the latent variable. The psychometric properties of the PI under this more realistic scenario will be studied in upcoming work.

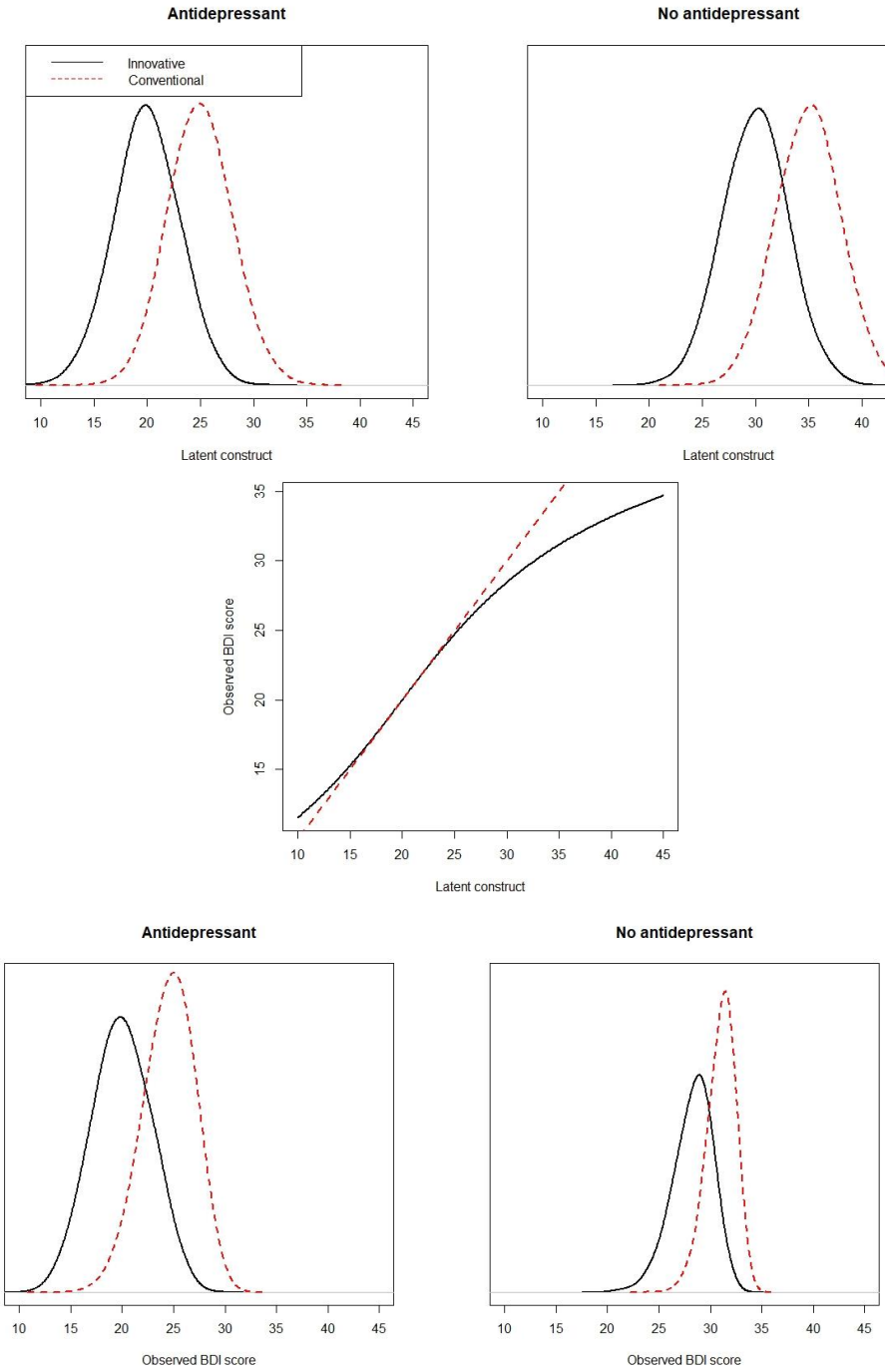


Figure 3. Artificial data where the innovative treatment outperforms the conventional treatment in terms of latent construct (lower = better) for both patients that receive antidepressants (left panels) or that do not receive antidepressants (right panels).The bottom panels show the results based on the observed BDI score, which can be obtained from the upper panels via the monotone transformation in the middle panel. For the upper panels there is no interaction in terms of the difference in means, while for the lower panels there is an interaction. The interaction in terms of the PI is the same for the latent construct as for the observed BDI scores.

2.3. Extension to a regression context

In this section, we discuss how Probabilistic Index Models (PIMs) can be used to extend the estimation of the PI in a regression context. For didactic reasons, we start with a univariate covariate. Once univariate PIMs are introduced, they can be easily extended to deal with multiple covariates.

2.3.1 One covariate

Consider a sample of n i.i.d. observations (Y_i, X_i). A PIM models the conditional PI directly as a function of the covariates. More specifically, a PIM is given by

$$\mathrm{P}\,(Y_i < Y_j | X_i, X_j) = \mathrm{m}(X_i, X_j; \beta). \quad (3)$$

Here $m(\cdot)$ is a user-specified function and it is the regression parameter that we want to estimate. As will be explained in Section 4, the following choice of $m(\cdot)$ will be convenient for a variety of applications:

$$\mathrm{m}\big(X_i, X_j; \beta\big) = g^{-1}[(X_j - X_i)\beta], \quad (4)$$

where $g(\cdot)$ denotes a link function mapping the unit-interval onto the real line, e.g. the logit or the probit link function. The interpretation follows from combining (3) and (4)

$$g^{-1}(\beta) = \mathrm{P}(Y_i < Y_j | X_i = x, X_j = x + 1). \quad (5)$$

Hence $g^{-1}(\beta)$ gives the probability that a randomly selected subject with covariate value x will have a lower outcome as compared to a randomly selected subject with a covariate value that is one unit higher. When X is binary and $x = 0$, then (5) gives the two-sample PI. The advantage of modeling the PI directly as a function of covariates is that (3) can also be used when X is continuous, where (5) gives the change in PI when X is increased by one unit. This is similar to the interpretation of a conventional linear regression model. When $\mathrm{E}(Y_i|X_i) = \alpha_0 + \alpha X_i$, then

$$\alpha = \mathrm{E}\big(Y_j \big| X_j = x + 1\big) - \mathrm{E}(Y_i | X_i = x) = \mathrm{E}\big(Y_j - Y_i \big| X_i = x, X_j = x + 1\big).$$

Whereas a PIM quantifies the effects in terms of an ordering between two outcomes, the linear regression models quantifies the effects in terms of the expected differences between two outcomes.

When X is binary, we know that we can estimate the PI via (2). How can we estimate the PI when X is continuous? The solution lies in rewriting the PI as an expectation. It holds that

$$\mathrm{P}\,(Y_i < Y_j |\, X_i,\, X_j) = \mathrm{E}\big(\,\mathrm{I}\,(Y_i < Y_j) |\, X_i,\, X_j\big)\,, \quad (6)$$

where I(·) is the indicator as defined in (2). We introduce the compact notation $I_{ij} = \mathrm{I}(Y_i < Y_j)$ and $X_{ij} = X_j - X_i$. Consider, for example, the logit-link $g(x) = \log[x/(1-x)]$ (in Section 4.2 we illustrate why this choice makes sense). Combing (3), (4) and (6) gives

$$\mathrm{E}\big(I_{ij}\big|X_{ij}\big) = \mathrm{expit}(X_{ij}\beta), \quad \mathrm{expit}(x) = \frac{e^x}{1+e^x}.$$

This is exactly a logistic regression model applied to the transformed binary outcomes I_{ij} and the transformed predictors $X_{ij}, (i, j = 1, \ldots, n)$. Thas et al. (2012) show that this strategy results in an asymptotically normal and consistent estimator of β. Hence fitting a PIM on the data $(Y_i,\, X_i)$ is equivalent to fitting a binary regression model to the transformed data $(I_{ij},\, X_{ij})$. By rewriting a PIM as a binary regression model we can use existing software to fit PIMs in practice, e.g. **glm** in R (R Core Team, 2017). Therefore an estimator $\hat{\beta}$ for β in (3) is obtained by solving the quasi likelihood estimating equations

$$\sum_{i=1}^{n}\sum_{j=1}^{n} A\big(X_i, X_j; \beta\big)\big[I_{ij} - m\big(X_i X_j; \beta\big)\big] = 0,$$

$$A\big(X_i, X_j; \beta\big) = \frac{\frac{\partial m(X_i X_j;\beta)}{\partial \beta}}{m(X_i X_j;\beta)[1-m(X_i X_j;\beta)]}\,. \quad (7)$$

However, when relying on standard software that assumes i.i.d. data, we cannot trust the standard errors. Despite the data $(Y_i,\, X_i)$ being i.i.d., the transformed data $(I_{ij},\, X_{ij})$ are no longer independent. To illustrate this, consider two transformed outcomes $I_{ij} = \mathrm{I}(Y_i < Y_j)$ and $I_{ik} = \mathrm{I}(Y_i < Y_k)$. Both binary outcomes share Y_i, making them no longer independent. The transformed outcomes I_{ij} have a cross-correlation structure, which is different from the typical block correlation structure in multilevel or longitudinal data. Hence, standard errors that assume independent or multilevel data will not be consistent estimators of the standard errors associated with a PIM. Thas et al. (2012) therefore provide a consistent sandwich estimator for the standard errors that takes the cross-correlation into account. These standard errors are implemented in the R package **pim** (Meys et al., 2017). Note that rewriting a PIM as a binary regression model has implications for the computational complexity: whereas the original sample is of size n, the transformed sample $(I_{ij},\, X_{ij})$ has n^2 elements (of which $n(n-1)/2$ are redundant). This increases the computational complexity drastically.

In summary, the theory provided by Thas et al. (2012) and implemented in the **pim** package provides a consistent estimator for β that is asymptotically normal and a consistent estimator for its standard error ($SE_{\widehat{\beta}}$). It is then straightforward to construct $(1-\alpha)100\%$ confidence intervals for β via $(\widehat{\beta} - z_{\alpha/2}SE_{\widehat{\beta}}, \widehat{\beta} + z_{\alpha/2}SE_{\widehat{\beta}})$, with $z_{\alpha/2}$ the quantile so that $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. Because $g^{-1}(\cdot)$ is strictly increasing, a confidence interval for the PI (i.e. $g^{-1}(\beta)$) can be obtained by transforming the boundaries of this interval: $(g^{-1}[\widehat{\beta} - z_{\alpha/2}SE_{\widehat{\beta}}], g^{-1}[\widehat{\beta} + z_{\alpha/2}SE_{\widehat{\beta}}])$. Hypothesis tests for $\beta = \beta_0$ are obtained by constructing Wald-statistics $(\hat{\beta} - \beta_0)/SE_{\widehat{\beta}}$ which have an asymptotic standard normal null distribution. Since Wald-statistics can perform poorly in small samples, Amorim et al. (2017) proposed a bias-reduced version of the bootstrap and adjusted jackknife empirical likelihood that lead to drastic improvements in small sample inference for PIMs. Discussing these estimators, however, falls beyond the scope of this article.

Despite the close connection with logistic regression, it is important to mention that the link-function of a PIM plays a different role than the link function in binary regression. In logistic regression, we model the probability of a 'success' as a function of the covariate via $\mathrm{P}(\mathrm{succes}|X_i) = \mathrm{expit}(\gamma_0 + \gamma X_i)$. From this model, we then derive an interpretation of γ via

$$\exp(\gamma) = \frac{\mathrm{odds}(\mathrm{succes}|X_i=x+1)}{\mathrm{odds}(\mathrm{succes}|X_i=x)}, \qquad \mathrm{odds}(\mathrm{succes}|X_i) = \frac{\mathrm{P}(\mathrm{succes}|X_i)}{1-\mathrm{P}(\mathrm{succes}|X_i)}.$$

Here the choice of link function is crucial in obtaining this odds ratio interpretation. When a different link function is used, e.g. the probit, then $\exp(\gamma)$ has no interpretation in terms of an odds ratio. For a PIM, this is different. From (5), we see that we can always transform β via $g^{-1}(\cdot)$ to get an interpretation in terms of the PI. This holds for all link functions (logit, probit, cloglog, etc.). This is a consequence of the fact that a PIM models an effect size directly, whereas in logistic regression the effect sizes are derived from modeling the probability of success.

2.3.2 Multiple covariates

The rationale of Section 3.1 can be easily adopted to a multi-variable context. Let $\boldsymbol{X}$ denote the *p*-dimensional vector of covariates associated with *Y*, then a PIM is given by

$$\mathrm{P}\,(Y_i < Y_j \,|\, \boldsymbol{X}_i, \boldsymbol{X}_j) = g^{-1}[(\boldsymbol{X}_j - \boldsymbol{X}_i)^T \beta], \qquad (8)$$

with β the *p*-dimensional vector of interest. To estimate β we transform the outcomes as before to I_{ij} and similarly to the vector of covariates $\boldsymbol{X}_{ij}^T = \boldsymbol{X}_j - \boldsymbol{X}_i$. Similar to how logistic regression can deal with multiple covariates, so can the PIM. The estimator provided by Thas et al. (2012), say $\hat{\beta}$, will be consistent and asymptotically multivariate normal. Their sandwich estimator will consistently estimate the variance-covariance matrix of $\hat{\beta}$. This allows for construction of quadratic test statistics to test multivariate hypotheses and to construct multivariate confidence regions.

To illustrate the interpretation, we consider a bivariate regressor: $\boldsymbol{X}^T = (Z_1, Z_2)$ with $\beta^T = (\beta_1, \beta_2)$. From (8), it follows that

$$g^{-1}(\beta_1) = \mathrm{P}(Y_i < Y_j | Z_{1i} = z, Z_{1j} = \mathrm{z} + 1, Z_{2i} = Z_{2j}),$$

i.e. $g^{-1}(\beta_1)$ gives the probability that a randomly selected subject with covariate value *z* for Z_1 will have a lower outcome as compared to a randomly selected subject with a covariate value that is one unit higher, *where Z_2 is the same for both subjects*. PIMs can thus be used to estimate the PI associated with a unit increase in one covariate, while controlling for the other covariates.

Note that PIM (8) has no intercept. Indeed, if $\boldsymbol{X}_i = \boldsymbol{X}_j$ then it must follow that $\mathrm{P}\left(Y_i < Y_j \middle| \boldsymbol{X}_i, \boldsymbol{X}_j\right) = 0.5$ because Y_i and Y_j have the same conditional distribution. If $g^{-1}(\cdot)$ denotes a conventional link function such as the logit or profit, model (9) implies that $\mathrm{P}\left(Y_i < Y_j \middle| \boldsymbol{X}_i, \boldsymbol{X}_j\right) = g^{-1}(0) = 0.5$ and hence the intercept has to be equal to zero.

2.3.3 Dealing with ties

So far we have ignored discrete outcomes or outcomes with ties. To accommodate for this, we extend the definition of the PI to $\mathrm{P}\left(Y_i < Y_j \middle| \boldsymbol{X}_i, \boldsymbol{X}_j\right) + \frac{1}{2}\mathrm{P}\left(Y_i = Y_j \middle| \boldsymbol{X}_i, \boldsymbol{X}_j\right)$. This PIM estimation theory can now be adopted by considering the transformed outcomes $\mathrm{I}(Y_i < Y_j) + 0.5\,\mathrm{I}(Y_i = Y_j)$. The remainder of the estimation theory is unaffected by these ties. We refer to Thas et al. (2012) for more details.

2.4. Comparison with other methods

For a better understanding of PIMs, we study the relationship with several other well-known methods. We start with the parametric regression model with normal errors and then extend this to a semiparametric context. The connection with the Wilcoxon–Mann–Whitney test is also discussed.

2.4.1 The normal linear model

The linear regression model with normal errors is given by

$$Y_i = \alpha_0 + \boldsymbol{X}_i^T \boldsymbol{\alpha} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2). \quad (9)$$

When we plug in this model in the PI, it follows that

$$\mathrm{P}\left(Y_i < Y_j \middle| \boldsymbol{X}_i, \boldsymbol{X}_j\right) = \mathrm{P}\left(\epsilon_i - \epsilon_j < (\boldsymbol{X}_j - \boldsymbol{X}_i)^T \boldsymbol{\alpha} \middle| \boldsymbol{X}_i, \boldsymbol{X}_j\right) = \Phi\left((\boldsymbol{X}_j - \boldsymbol{X}_i)^T \frac{\boldsymbol{\alpha}}{\sqrt{2}\sigma}\right), \quad (10)$$

where the last step comes from the fact that $\epsilon_i - \epsilon_j$ follows a normal distribution with a mean of zero and variance of $2\sigma^2$. From this relationship we recognize the form (8) with probit-link $g^{-1}(\cdot) = \Phi(\cdot)$ and $\beta = \boldsymbol{\alpha}/\sqrt{2}\sigma$.

The relationship (11) further implies that we can estimate β in (8) using the maximum likelihood estimators of α and σ under model (9), i.e. $\hat{\beta} = \hat{\boldsymbol{\alpha}}/\sqrt{2}\hat{\sigma}$. This approach is discussed in Tian (2008). This estimator will, however, only be consistent if the parametric model (9) holds, while the estimator proposed by Thas et al. (2012) will be consistent under a broader class of data generating models. This is demonstrated in the following section.

2.4.2 Semiparametric linear transformation models

The parametric linear model can be extended to a semiparametric context by introducing a nonparametric component. We start by considering the semiparametric linear transformation model (SLTM) with normal error:

$$h(Y_i) = \alpha_0 + \boldsymbol{X}_i^T\boldsymbol{\alpha} + \epsilon_i\ ,\ \epsilon_i \sim N(0,\sigma^2)\ (11)$$

where $h(\cdot)$ denotes an *unknown* strict monotone function which reflects the nonparametric component. The distribution of ϵ_i is fully specified (typically by setting $\sigma^2 = 1$) and reflects the parametric component of the model. The nonparametric and the parametric parts give rise to a semiparametric model. The Box-Cox transformation model (Box & Cox, 1964) is a special case of (11) where the function $h(\cdot)$ belongs to a parametric family. We refer to Cheng, Wei, Ying (1995); Chen, Jin, Ying (2002); Zeng and Lin (2007) and the references therein for detailed discussions on SLTMs.

Since the PI is monotone invariant, see Section 2.4, the algebra of Section 4.1 can be repeated to obtain the same relationship (11). This immediately demonstrates that the PIM is a genuine semiparametric model. Hence, the PIM (8) with probit-link holds from the moment one can find a monotone transformation so that the linear model with normal error holds when the transformed outcome is modeled. This relationship can be exploited to assess the goodness-of-fit of PIMs. We demonstrate this in Section 5.

Instead of the normal error, there is another choice of error distribution that gives rise to a very popular semiparametric model. More specially, we consider the SLTM

$$h(Y_i) = \alpha_0 + \boldsymbol{X}_i^T\boldsymbol{\alpha} + \epsilon_i,\ \epsilon_i \sim F_{EV}(e) = 1 - \exp(-\exp[e]),\ (12)$$

where $F_{EV}(\cdot)$denotes the extreme value distribution. Using basic algebra, model (12) can be equivalently written as

$$\log[-\log \mathrm{P}(Y_i > y|\boldsymbol{X}_i)] = h^*(y) - \boldsymbol{X}_i^T\boldsymbol{\alpha},\ h^*(y) := h(y) - \alpha_0.\ (13)$$

This is the famous Cox proportional hazards model (Cox, 1972). To establish the connection with a PIM we substitute the SLTM in the PI:

$$\begin{aligned}\mathrm{P}\left(Y_i < Y_j \middle| \boldsymbol{X}_i, \boldsymbol{X}_j\right) &= \mathrm{P}\left(h(Y_i) < h(Y_j) \middle| \boldsymbol{X}_i, \boldsymbol{X}_j\right)\\ &= \mathrm{P}\left(\epsilon_i - \epsilon_j < [\boldsymbol{X}_j - \boldsymbol{X}_i]^T\boldsymbol{\alpha}\right)\\ &= \mathrm{expit}[(\boldsymbol{X}_j - \boldsymbol{X}_i)^T\boldsymbol{\alpha}].\end{aligned}$$

The last step follows form the fact that the difference $\epsilon_i - \epsilon_j$ follows a standard logistic distribution if ϵ_i and ϵ_j follow the extreme value distribution. This implies that the Cox proportional hazards model (12) gives rise to a PIM of the form (8) with logit-link and $\boldsymbol{\beta} = \boldsymbol{\alpha}$.

It is worth mentioning that there is no one-to-one correspondence between PIMs and SLTMs, since multiple SLTMs can give rise to the same PIM. For example, the SLTMs with product-normal error and the SLTMs with exponential error both give rise to a PIM of the form (12) with a Laplace link.

2.4.3 Rank tests

In De Neve and Thas (2015), it is shown how different rank tests can be obtained by formulating an appropriate PIM. Here we focus on the two-sample design. We consider the context and notation of Section 2.3 and let $X_i = 0$ for the innovative treatment and $X_i = 1$ for the conventional treatment. Before we focus on the PIM, we first consider the linear regression model to set the scene. To study the association between X and Y we can fit the linear regression model

$$E(Y_i|X_i) = \alpha_0 + \alpha X_i$$

Testing H_0: $\alpha = 0$ using least squares is now equivalent to performing a two-sample *t*-test. The advantage of writing the problem as a regression model is that the latter can be used to study associations when e.g. confounders are present. These confounders, say $\boldsymbol{Z}_i$, can be controlled for by including them as predictors in the regression model

$$E(Y_i|X_i, \boldsymbol{Z}_i) = \alpha_0 + \alpha X_i + \boldsymbol{\gamma}^T \boldsymbol{Z}_i,$$

and testing H_0: $\alpha = 0$ now accounts for the confounders since

$$\alpha = E(Y_i|X_i = 1, \boldsymbol{Z}_i = \boldsymbol{z}) - E(Y_i|X_i = 0, \boldsymbol{Z}_i = \boldsymbol{z})$$

Suppose now that instead of using the *t*-test we would like to use a rank test to study the association between Y and X. For the two-sample design, the Wilcoxon–Mann–Whitney test acts as the rank-counterpart of the two-sample *t*-test. If we were able to embed this rank test in a regression model, then we could, similarly as in the linear regression model, extend the rank test by including confounders. It turns out that PIMs are perfectly suited for this.

We consider the PIM (3) with model formulation (4). The estimator of β solves the estimating equation (78) and it is not difficult to show that $\hat{\beta} = g(MW)$.

Hence, the estimator of the PIM parameter β is related to the Mann–Whitney statistic (2). Moreover, De Neve and Thas (2015) demonstrate how a score test for $H_0: \beta = 0$ can be constructed that is equivalent to the WMW test. Hence, we can extend this test to deal with confounders by reformulating it as a PIM:

$$\mathrm{P}\left(Y_i < Y_j \middle| X_i, X_j, \boldsymbol{Z}_i, \boldsymbol{Z}_j\right) = g^{-1}[\beta\left(X_j - X_i\right) + (\boldsymbol{Z}_j - \boldsymbol{Z}_i)^T \boldsymbol{\gamma}].$$

The null hypothesis $H_0: \beta = 0$ is equivalent to $\mathrm{P}\left(Y_i < Y_j \middle| X_i = 0, X_j = 1, \boldsymbol{Z}_i = \boldsymbol{Z}_j\right) = 0.5$. This demonstrates that we are comparing subjects from different treatment groups, but with the same value of the confounders. A score- or Wald-test can then be constructed to test $H_0: \beta = 0$ and a confidence interval for β can be constructed.

Similar connections can be established for other rank tests, such as the Kruskal–Wallis rank test and the Friedman rank test; we refer to De Neve and Thas (2015) for details. In summary, the PIM allows one to embed rank tests into a regression framework and this has the following two main advantages: 1) the rank tests can be extended to more complicated designs and 2) in addition to hypothesis testing, effect sizes can be derived and confidence intervals can be constructed.

2.5. Goodness-of-fit

Because PIMs are semiparametric, inference is only valid when the proposed model is consistent with the underlying data-generating model. Therefore, it is important to assess the goodness-of-fit (GOF) of a PIM. De Neve et al. (2013a) developed a formal GOF-test together with GOF-plots for PIMs. These methods rely on nonparametric smoothers and hence suffer from the curse of dimensionality. This makes them not useful in many practical situations. We will therefore address the GOF differently. More specifically, we will exploit the connection with the SLTMs of Section 4.2. We consider the following procedure:

1. Fit model (9) to the data.
2. Check the assumptions of the linear model (linearity of the model and constant variance and normality of the errors). If they are fulfilled, go to step 3. If some of the assumptions are violated, go to step 4.

3. Due to the connection with normal linear model, PIM (10) will be consistent with the underlying data-generating model.
4. Perform a Box–Cox transformation on the linear regression model.
5. Check the assumptions of the linear model applied to the Box–Cox transformed outcome (linearity of the model and constant variance and normality of the errors). If they are fulfilled, go to step 6. If some of the assumptions are violated, go to step 7.
6. Due to the connection with SLTM with normal error, PIM (10) will be consistent with the underlying data-generating model.
7. This is the most difficult scenario: a SLTM does not fit the data and hence we cannot exploit its connection with PIMs to assess the GOF. Once an appropriate (possibly non-linear) regression model is established, its connection with the PIM will have to be worked out to assess the GOF.

The above procedure is heuristic and it is clear that Step 7 is not very useful in practice. This indicates that more research on assessing the adequacy of PIM is needed. Note that if a Box–Cox transformation makes the model linear and the variance of the error constant, but the residuals do not exhibit normality a PIM with a different link-function can be used. If e.g. the residuals exhibit a skewed distribution that can be approximated by the extreme-value distribution, then a PIM with logit-link is appropriate; see Section 4.2.

2.6. Illustration

We illustrate the methodology on data from a clinical trial to evaluate a computerized, interactive cognitive behavioral therapy for patients with depression. The original study is reported in Proudfoot et al. (2003) and part of the data are available in the R package of Hothorn and Everitt (2017a,b). Patients with depression were recruited in primary care and were randomized over two treatments: an innovative treatment or a conventional treatment. The conventional treatment (TAU: treatment as usual) consisted of a face-to-face cognitive behavioral therapy, while the innovative treatment consisted of an interactive computerized program called Beat the Blues™ (BtheB) replacing the face-to-face counseling. We refer to Proudfoot et al. (2003) for details. The case study only serves as an illustration for PIMs and substantive conclusions of the study can be found in Proudfoot et al. (2003). We include the

most important R code in the text to illustrate how PIMs can be used in practice. An R-markdown document with all R code can be found in the supplementary material (Appendix 1).

We consider the following variables: the Beck Depression Inventory II score at baseline (`bdi.pre`) and after three months (`bdi.3m`), the treatment (`treatment`: TAU or BtheB) and weather the patient takes anti-depressant drugs (`drug`: yes or no). It is of interest to study 1) the association between the treatment and the depression score and 2) the association between antidepressants and the depression score.

Figure 4 (top left panel) gives the boxplots with strip charts of the depression scores after three months for both treatments. The two-sample *t*-test indicates a marginal significant effect ($p = 0.041$) in favor of the innovative treatment.

```
> t.test(bdi.3m~treatment, var.equal=TRUE, data=Data)

Two Sample t-test

data: bdi.3m by treatment
t = 2.0849, df = 71, p-value = 0.04067
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.2461034 11.0331759
sample estimates:
mean in group TAU mean in group BtheB
17.66667 12.02703
```

As discussed in Section 4, we can obtain similar results when using a linear regression model:

```
> fit.lm <- lm(bdi.3m~treatment, data=Data)
> coef(summary(fit.lm))
                  Estimate    Std. Error  t value     Pr(>|t|)
(Intercept)       17.66667    1.925751    9.173908    1.121148e-13
treatmentBtheB    -5.63964    2.704960    -2.084926   4.067314e-02
```

The boxplot reveals that there is an outlier in the innovative treatment group. When we repeat the analysis without this outlier, the *p*-value decreases to 0.008. Since the result of the *t*-test depends heavily on the value of this patient, it might be desirable to use the Wilcoxon–Mann–Whitney test because rank procedures are more robust to outliers.

```
> wilcox.test(bdi.3m~treatment,exact=FALSE, data=Data)

Wilcoxon rank sum test with continuity correction

data: bdi.3m by treatment
W = 851.5, p-value = 0.04104
alternative hypothesis: true location shift is not equal to 0
```

The test statistic $W = 851.5$ equals $\sum_{i,j} I(Y_{TAU,i} > Y_{BtheB,j})$ so that the PI is estimated by

```
> n1 <- length(which(Data$treatment=="TAU"))
> n2 <- length(which(Data$treatment=="BtheB"))
> W <- wilcox.test(bdi.3m~treatment,exact=FALSE, data=Data)$statistic
> PI <- W/(n1*n2)
> PI
W
0.6392643
```

Hence, $\widehat{\mathrm{P}}(Y_{TAU} > Y_{BtheB}) = 0.64$, or equivalently, $\widehat{\mathrm{P}}(Y_{TAU} < Y_{BtheB}) = 0.36$. The estimated probability that a patient receiving BtheB will have a large depression score as compared to a patient receiving TAU is 36%. It is therefore less likely that patients receiving BtheB will have higher BDI scores. As discussed in Section 4 we can obtain similar results by fitting a PIM. In R, this can be done via the function `pim()` in the R-package **pim**. The interface of `pim()` is similar to that of `lm()` or `glm()`:

```
> fit.pim <- pim(bdi.3m~treatment,link="probit", data=Data)
> summary(fit.pim)
pim.summary of following model :
bdi.3m ~ treatment
Type: difference
Link: probit

                 Estimate    Std. Error  z value     Pr(>|z|)
treatmentBtheB    -0.3565     0.1703      -2.093      0.0363 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Null hypothesis: b = 0
```

Note that the PIM *p*-value slightly deviates from the *p*-value of the WMW test, since `pim()` constructs Wald statistics by default, while the WMW test uses a score statistic. We can now derive the interpretation of this coefficient from (6). It follows that

$$0.36 = \Phi(-0.3565) = \widehat{\mathrm{P}}\left(Y_i < Y_j | X_i = TAU, X_j = BtheB\right) = \widehat{\mathrm{P}}(Y_{TAU} < Y_{BtheB}).$$

Because of equation (15), this is exactly the same estimate as provided by the WMW test. A nice feature of reformulating the WMW test as a PIM is that we can easily obtain confidence intervals (CI) for the effect size via the `confint()` function:

```
> confint(fit.pim)
                   2.5 %       97.5 %
treatmentBtheB    -0.6902768  -0.02270898
> pnorm(confint(fit.pim))
                   2.5 %       97.5 %
treatmentBtheB    0.2450101   0.4909412
```

The first R code gives the 95% CI for β, while the second for $\Phi(\beta)$ gives the interpretation on the PI scale. The 95% CI for $\mathrm{P}\left(Y_i < Y_j | X_i = TAU, X_j = BtheB\right)$ goes from 0.25 to 0.49. Note that this interval is quite wide and the upper bound is close to 50%, providing only marginal evidence that BtheB is superior as compared to TAU in terms of the BDI score. This is also reflected by the p-value which is close to 5%.

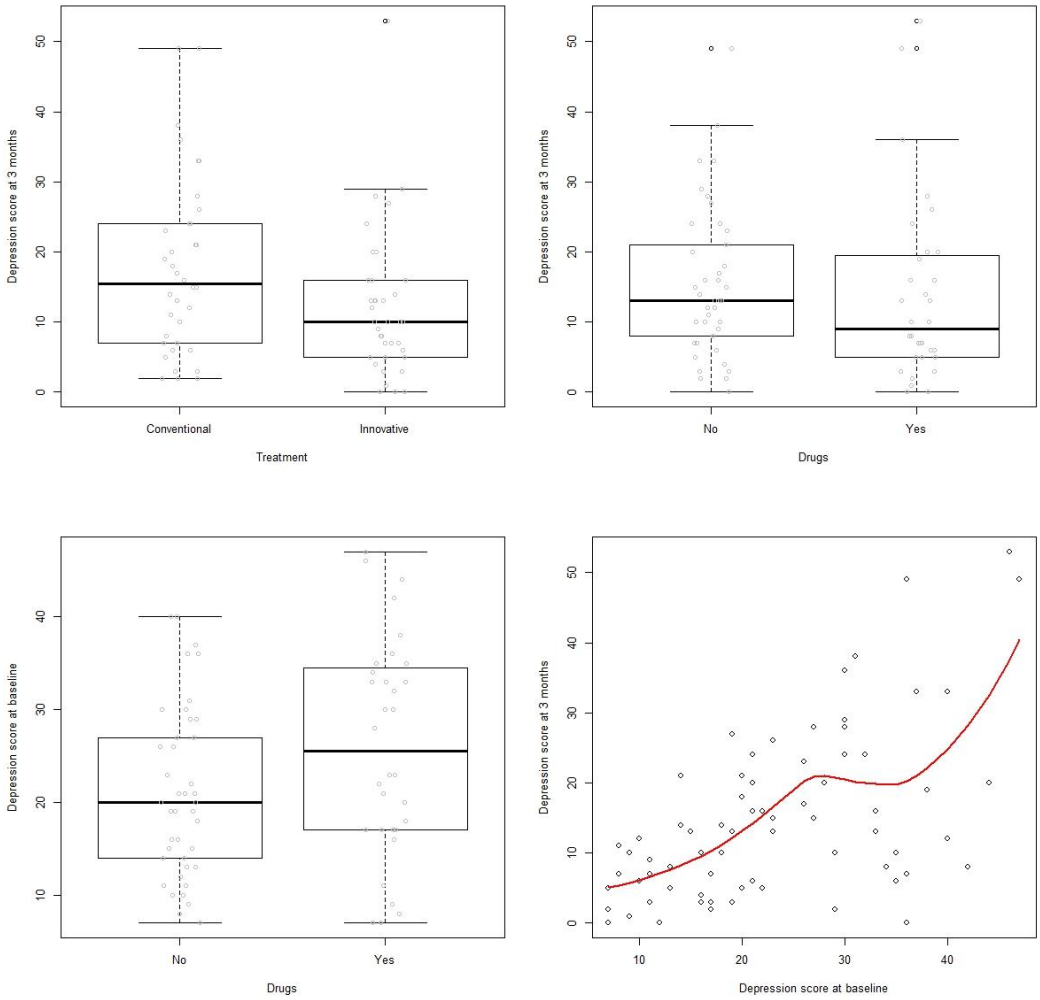


Figure 4. BDI scores at 3 months according to treatment group (top left) or antidepressant (top right). BDI scores at baseline according to antidepressants (bottom left) and BDI scores at three months as a function of the BDI scores at baseline (bottom right) with a smoother.

We now study the association between antidepressants and the depression score. From the top right panel of Figure 4 we see a slightly reduced depression score for patients receiving antidepressants. Based on the WMW test, the PI is not significantly different from 50% (p = 0.20).

```
> wilcox.test(bdi.3m~drug,exact=FALSE, data=Data)
```

```
Wilcoxon rank sum test with continuity correction

data: bdi.3m by drug
W = 772, p-value = 0.1987
alternative hypothesis: true location shift is not equal to 0
```

As before, a similar result is obtained when we fit a PIM:

```
> fit.pim2 <- pim(bdi.3m~drug,link="probit", data=Data)
> summary(fit.pim2)
pim.summary of following model :
bdi.3m ~ drug
Type: difference
Link: probit

            Estimate    Std. Error  z value     Pr(>|z|)
drugYes     -0.2235     0.1718      -1.3        0.193
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Null hypothesis: b = 0
```

Since antidepressants were not randomized over patients we have to be careful in interpreting the association. From the bottom panels of Figure 4 it is clear that the depression score at baseline acts as a confounder. We can account for this confounder by including it as a covariate in a PIM. In R, this becomes:

```
> fit.pim3 <- pim(bdi.3m~drug+bdi.pre,link="probit", data=Data)

> summary(fit.pim3)
pim.summary of following model :
bdi.3m ~ drug + bdi.pre
Type: difference
Link: probit


Estimate Std. Error z value Pr(>|z|)
drugYes -0.522230 0.185775 -2.811 0.00494 **
bdi.pre 0.049163 0.009856 4.988 6.09e-07 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null hypothesis: b = 0
```

When we account for the baseline depression score, the effect of antidepressants becomes significant (p = 0.005). The estimated PI equals $\widehat{\mathrm{P}}\left(Y_i < Y_j | X_{drugYes,i} = 0, X_{drugYes,j} =\right.$

$1, X_{BDIpre,i} = 0, X_{BDIpre,j}\big) = \Phi(-0.522230) = 30\%$ (95% CI 19% - 44%). When we compare two patients with the same BDI score at baseline, the probability that the BDI score at 3 months is lower for the patient that does not take antidepressant is estimated as 30%. It is therefore unlikely that not taking antidepressants is associated with lower BDI scores, while controlling for the baseline BDI.

A PIM also allows quantification of the association between two continuous variable. Form the PIM output it follows that
$\widehat{\mathrm{P}}\big(Y_i < Y_j | X_{drugYes,i} = X_{drugYes,j}, X_{BDIpre,j} = X_{BDIpre,i} + 10\big) = \Phi(10 \times 0.049163) = 69\%$. For two patients with the same antidepressant status (both receive antidepressant or both receive no antidepressant), the probability that the patient that has a BDI score at baseline that is 10 units higher will have the highest BDI score after 3 months is estimated as at 69%. It is therefore more likely that patients with higher BDI scores at baseline will be associated with higher scores after three months.

Figure 5 (left panel) displays the residual plot of the linear model `lm(bdi.3m ~ drug + bdi.pre)`. The plot indicates that the residual variance is not constant. Because the presence of non-positive values in our data, we opt for a two-parameter Box–Cox transformation. In contrast to the more common (one-parameter) Box–Cox transformation, a shift parameter (λ_2) is additionally estimated. Transforming the outcome via $h(Y) = \frac{(Y+\lambda_2)^{\lambda_1}-1}{\lambda_1}$, with $\lambda_1 = 0.46759$ and $\lambda_2 = 0.00053$ (the obtained values of the two-parameter Box–Cox transformation) stabilizes this variance and makes the residuals approximately normal; see Figure 5 (middle and right panel). This implies that the PIM with probit link can be used as an approximation of the data-generating model.

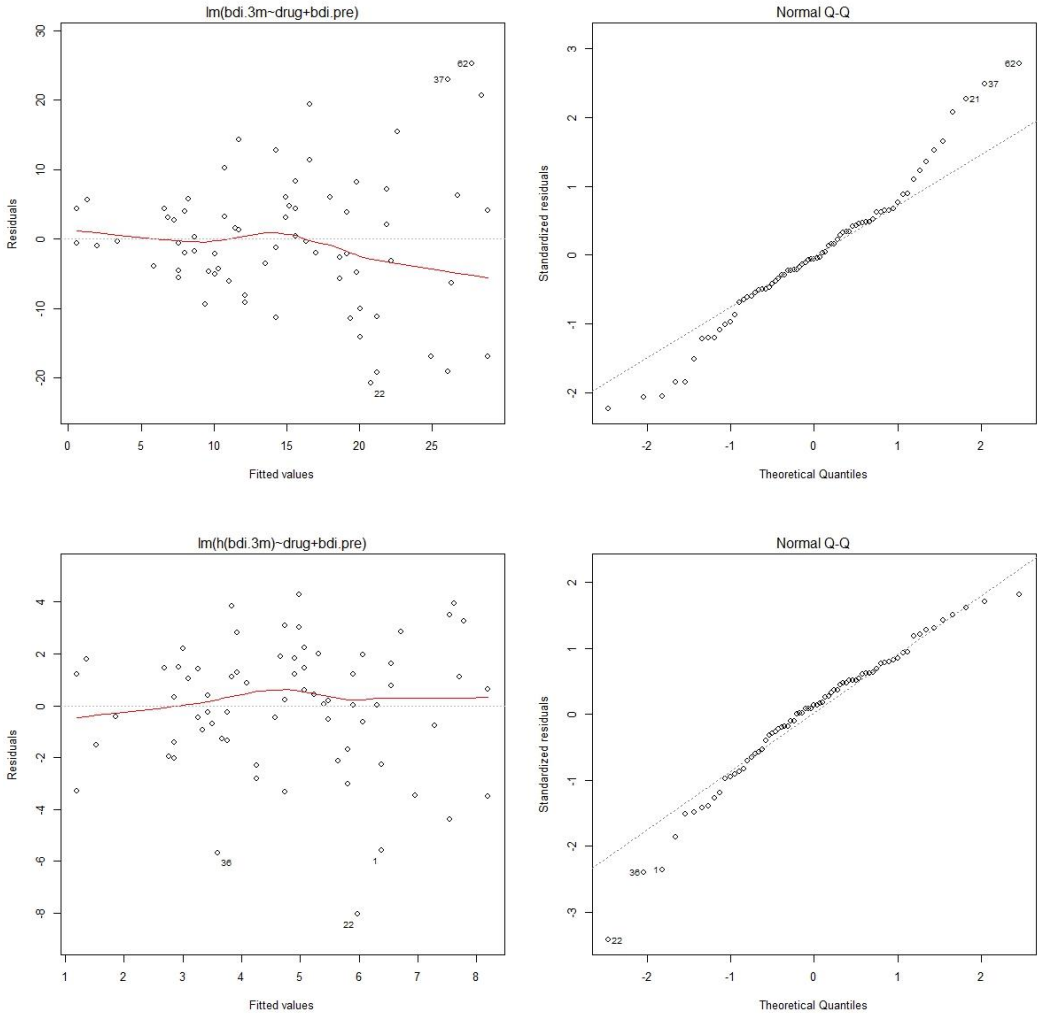


Figure 5. Left panel: residual plot of the linear model. Middle and right panel: residual and QQ-plot of the linear model when the outcome is transformed according to a Box–Cox transformation h.

2.7 Conclusions

This article serves as an introduction to PIMs allowing summarizing the association between an outcome and a covariate in terms of the PI. This is fundamentally different from conventional regression models where associations are expressed in terms of mean differences. A substantial part of this tutorial is devoted to the properties of the PI as an effect size. This is essential because of good understanding of PIMs starts with a good understanding of the PI. Via this tutorial, we hope to stimulate the community to apply and investigate these models in the behavioral sciences. Despite several publications on PIMs and the availability of the R package **pim**, more research has to be conducted to make these models applicable for many practical situations. Challenges include, among others, 1) the extension of PIMs to multilevel data 2) embedding latent variables in PIMs and 3) extending goodness-of-fit methods.

References

Acion, L., Peterson, J., Temple, S., and Arndt, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*, 25:591–602.

Amorim, G., Thas, O., Vermeulen, K., Vansteelandt, S., and De Neve, J. (2017). Small sample inference for probabilistic Index models. *Computational Statistics and Data Analysis*, under review.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations (with Discussion). *Journal of the Royal Statistical Society. Series B*, pages 211–252.

Brumback, L. C., Pepe, M. S., and Alonzo, T. A. (2006). Using the ROC curve for gauging treatment effect in clinical trials. *Statistics in Medicine*, 25(4):575–590.

Brunner, E. and Munzel, U. (2000). The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. *Biometrical journal*, 42(1):17–25.

Chen, K., Jin, Z., and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, 89(3):659–668.

Cheng, S.,Wei, L., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, 82(4):835– 845.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494.

Cox, D. (1972). Regression models and life tables (with Discussion). Journal of the Royal *Statistical Society. Series B*, 34:187–220.

D'Agostino, R. B., Campbell, M., and Greenhouse, J. (2006). The Mann–Whitney statistic: continuous use and discovery. *Statistics in Medicine*, 25:541–542.17

Davison, M. L. and Sharma, A. R. (1990). Parametric statistics and levels of measurement: Factorial designs and multiple regression. *Psychological Bulletin*, 107(3):394.

De Neve, J., Meys, J., Ottoy, J.-P., Clement, L., and Thas, O. (2014). UnifiedWMWqPCR: the unified Wilcoxon–Mann–Whitney test for analyzing RT-qPCR data in R. *Bioinformatics*, 30(17):2494–2495.

De Neve, J. and Thas, O. (2015). A regression framework for rank tests based on the probabilistic index model. *Journal of the American Statistical Association*, 110(511):1276–1283.

De Neve, J., Thas, O., and Ottoy, J.-P. (2013a). Goodness-of-fit methods for probabilistic index models. *Communications in Statistics - Theory and Methods*, 42(7):1193–1207.

De Neve, J., Thas, O., Ottoy, J.-P., and Clement, L. (2013b). An extension of the Wilcoxon-Mann-Whitney test for analyzing RT-qPCR data. *Statistical Applications in Genetics and Molecular Biology*, 12(3):333–346.

Dodd, L. E. and Pepe, M. S. (2003). Semiparametric regression for the area under the receiver operating characteristic curve. *Journal of the American Statistical Association*, 98(462):409–417.

Garcia-Marques, L., Garcia-Marques, T., and Brauer, M. (2014). Buy three but get only two: The smallest effect in a 2x2 anova is always uninterpretable. *Psychonomic bulletin & review*, 21(6):1415–1430.

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*. 79(2):314-316.

Grissom, R. J. and Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods*, 6(2):135.

Grissom, R. J. and Kim, J. J. (2005). *Effect sizes for research: A broad practical approach.* Mahwah, NJ: Erlbaum.

Hauck, W. W., Hyslop, T., and Anderson, S. (2000). Generalized treatment effects for clinical trials. *Statistics in medicine*, 19(7):887–899.

Hothorn, T. and Everitt, B. S. (2017a). *A handbook of statistical analyses using R*. CRC press.

Hothorn, T. and Everitt, B. S. (2017b). HSAUR3: *A Handbook of Statistical Analyses Using R* (3rd Edition). R package version 1.0-6.

Kieser, M., Friede, T., and Gondan, M. (2013). Assessment of statistical significance and clinical relevance. *Statistics in Medicine*, 32:1707–1719.

Kotz, S. and Pensky, M. (2003). *The stress-strength model and its generalizations: theory and applications*. World Scientific.

Kruskal, W. H. (1952). A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics*, pages 525–540.

Laine, C. and Davidoff, F. (1996). Patient-centered medicine: a professional evolution. *Journal of the American Medical Association*, 275:152–156.

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *The Annals of Mathematical Statistics*, pages 165–179.

Lehmann, E. L. (2004). *Elements of large-sample theory*. Springer Science & Business Media.

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, 6(3):312–319.

Mann, H. and Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60.

Meys, J., De Neve, J., Sabbe, N., and Guimaraes de Castro Amorim, G. (2017). *pim: Fit Probabilistic Index Models*. R package version 2.0.1.

Newcombe, R. (2006). Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 1: general issues and tail-area-based methods. *Statistics in Medicine*, 25:543–557. 18

Proudfoot, J., Goldberg, D., Mann, A., Everitt, B., Marks, I., and Gray, J. (2003). Computerized, interactive, multimedia cognitive-behavioural program for anxiety and depression in general practice. *Psychological medicine*, 33(02):217–227.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruscio, J. (2008). A probability-based measure of effect size: robustness to base rates and other factors. *Psychological methods*, 13(1):19.

Ruscio, J. and Gera, B. L. (2013). Generalizations and extensions of the probability of superiority effect size estimator. *Multivariate behavioral research*, 48(2):208–219.

Ruscio, J. and Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research*, 47(2):201–223.

Senn, S. (1997). Letter to the editor: Testing for individual and population equivalence based on the proportion of similar responses, by d. m. rom and e. hwang, statistics in medicine, 15, 14891505 (1996). *Statistics in Medicine*, 16(11):1303–1305.

Senn, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects by L. Acion, J. Peterson, S. Temple and S. Arndt. *Statistics in Medicine*, 25:3944–3948.

Senn, S. (2011). U is for unease: reasons for mistrusting overlap measures for reporting clinical trials. *Statistics in Biopharmaceutical Research*, 3:302–309.

Thas, O. (2010). Comparing distributions. New York, NY: Springer.

Thas, O., De Neve, J., Clement, L., and Ottoy, J.P. (2012). Probabilistic index models (with Discussion). *Journal of the Royal Statistical Society - Series B*, 74:623–671.

Tian, L. (2008). Confidence intervals for P(Y1 > Y2) with normal outcomes in linear models. *Statistics in Medicine*, 27:4221–4237.

Vargha, A. and Delaney, H. D. (2000). A critique and improvement of the cl common language effect size statistics of mcgraw and wong. *Journal of Educational and Behavioral Statistics*, 25(2):101–132.

Vermeulen, K., Thas, O., and Vansteelandt, S. (2015). Increasing the power of the mann-whitney test in randomized experiments through flexible covariate adjustment. *Statistics in Medicine*, 34(6):1012–1030.

Wagenmakers, E.-J., Krypotos, A.-M., Criss, A. H., and Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after loftus. *Memory & cognition*, 40(2):145–160.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83.

Zeng, D. and Lin, D. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B* (Statistical Methodology), 69(4):507–564.

Zhou, W. (2008). Statistical inference for P(X < Y). *Statistics in Medicine*, 27:257–279. 19

Chapter 3

The Probabilistic Index:
A new effect size measure for the IAT

Maarten De Schryver, Helen Tibboel, Jan De Neve, Jan De Houwer
and Olivier Thas

Scoring algorithms are an important part of tests that are based on the Implicit Association Test (IAT) procedure. Besides describing which response latencies should be considered, they also define the effect size measure for estimating IAT-scores. At present, scoring algorithms based on D-effect size measures are used most often in IAT research. In the present paper, we introduce the Probabilistic Index (PI) as a candidate effect size measure for scoring data obtained from the IAT. Two new scoring algorithms are proposed using the PI-effect size measure. Compared to the D-effect size measure, the PI appears to be more robust against the influence of outliers. Moreover, PI-scoring algorithms outperform D-scoring algorithms on several psychometric criteria such as reliability and prediction.

3.1 Introduction

The Implicit Association Test (IAT) is a computer-based procedure that is widely used to measure implicit attitudes and stereotypes. Developed by Greenwald, McGhee and Schwartz (1998), the IAT aims to assess the relative strength of associations between two target (e.g., flowers and insects) and two attribute concepts (e.g., pleasant and unpleasant). The 'traditional' IAT[1] consists of seven blocks in which participants are instructed to classify target items and/or attribute items as quickly and accurately as possible (see Table 1 for an overview of the different phases of the IAT). During a first critical block of trials, items belonging to a first target (e.g., flowers) and attribute (e.g., pleasant) category are assigned to the same response key on a keyboard, while items belonging to a second target (e.g., insects) and attribute (e.g., unpleasant) category are assigned to another response key. In another critical block of trials, the response key assignments for the target categories are swapped, while the response key assignments for attributes categories remain the same.

The idea underlying the IAT is that the strength of mental associations in memory influences performance in the following manner: if the associations between the first target and first attribute and/or between the second target and second attribute stimuli are stronger than the associations between the first target and second attribute and/or second target and first attribute, then faster response latencies are expected in the first block than in the second block. If there are a priori reasons to believe that association strength does differ in this manner, then the first block is called the congruent block and the second block the incongruent block. The difference in performance between these two blocks (known as the IAT effect) is typically expressed as a single difference score. Although several interpretations of an IAT effect are possible, it is often regarded as a score reflecting the associative strength between a target and attributes relative to the associative strength between the second target and the same attributes (Greenwald, McGhee, & Schwartz, 1998).

[1] Different types of IAT have been developed such as the Brief IAT (Sriram & Greenwald, 2009) or Personalized IAT (Olson & Fazio, 2004). Although our arguments also apply to these other types of IAT, we will focus on the traditional IAT as outlined in this section.

The IAT has proved to be a valuable procedure in a wide range of contexts such as research on racial bias (e.g., Cunningham, Preacher, & Banaji, 2001), gender bias (e.g., Rudman, Greenwald, & McGhee, 2001), ageism (e.g., Levy & Banaji, 2002), and attitudes towards homosexuality (e.g., Banse, Seise, & Zerbes, 2001). Many researchers have used the IAT as a measure to predict behavior on an individual level in a wide variety of domains. For instance, researchers have tried to predict voting behavior on the basis of IAT scores (Arcuri, Castelli, Galdi, Zogmaister, & Amadori, 2008; Friese, Bluemke, & Wänke, 2007) and to predict consumers' preferences for different types of brands (Maison, Greenwald, & Bruin, 2004). Furthermore, researchers have used implicit measures like the IAT in a clinical context. They reasoned that implicit measures are especially useful to examine psychopathological behaviors because of their paradoxical nature: it is often obvious that such behaviors are counterproductive or irrational, but nevertheless people continue to perform them. It therefore seems that these behaviors are driven by processes that are implicit in the sense that they seem to be unintentional, uncontrollable, and maybe even unaware (e.g., Wiers, Teachman, & De Houwer, 2007).

Table 1. Overview of the seven phases of the classic IAT.

Phase	Block	Left Key	Right Key
Practice	B1	Target 1	Target 2
Practice	B2	Attribute pos	Attribute neg
Practice	B3	Target 1 + Attribute pos	Target 2 + Attribute neg
Test	B4	Target 1 + Attribute pos	Target 2 + Attribute neg
Practice	B5	Target 2 + Attribute pos	Target 1 + Attribute neg
Practice	B6	Target 2 + Attribute pos	Target 1 + Attribute neg
Test	B7	Target 2 + Attribute pos	Target 1 + Attribute neg

Note. pos = positive, neg = negative.

3.2 The IAT: Different scoring algorithms and effect size measures

Since the introduction of the IAT, different scoring algorithms have been proposed that provide an IAT-score. A typical *scoring algorithm* contains a set of consecutive steps one has to follow in order to determine an IAT-score for each participant. As outlined by Greenwald, Nosek and Banaji (2003), the proposed scoring algorithms specify the trials to be excluded, how the tails of the distribution and error-trials should be treated, which kind of transformation should be performed and how to calculate the IAT-score itself. We define *effect*

size measure as the equation used to calculate a quantity reflecting the magnitude of the difference in performance in the congruent and incongruent blocks. Because the effect size measure is only one aspect of the scoring algorithm, different scoring algorithms could involve similar or different effect size measures. Traditionally, the IAT-effect size measure was defined as the mean difference of the (log-transformed) response times observed in the incongruent and congruent blocks (referred to as the C-measure). An important limitation of this effect size measure (i.e., the mean difference, expressed in ms or log(ms)) is that these effects seem to correlate with general response speed (Fazio, 1990; Faust, Balota, Spieler & Ferraro,1999; Greenwald et al., 2003). This phenomenon causes biases such as larger effect sizes for elderly people who are typically slower than their younger counterparts. By using a standardized effect size measure (i.e., a measure that cancels out the unit of measurement; Kelley & Preacher, 2012), this artifact can be reduced, allowing researchers and practitioners to compare scores from different participants and/or groups, irrespective of their individual information processing rate (see Fazio, 1990; Faust et al., 1999).

Greenwald et al. (2003) introduced several D-scoring algorithms that involved a standardized effect size measure. Although these algorithms differ with respect to error-treatment and lower tail-treatment, they do share the same measure: the mean difference divided by the standard deviation of the pooled sample. Formally, this D-effect size measure can be expressed as:

$$D = \frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\frac{n_1 SD_1^2 + n_2 SD_2^2 + n_1 n_2 (\bar{Y}_1 + \bar{Y}_2)^2}{N}}} = \frac{r_{pb}}{\sqrt{\frac{n_1 n_2}{N^2 - N}}}$$

with $\bar{Y}_1$ ($\bar{Y}_2$) representing the sample mean of the response latencies observed in the congruent (incongruent) phase, SD_1^2 (SD_2^2) the squared standard deviation of response latencies observed in the congruent (incongruent) phase, n_1 (n_2) the number of trials in the congruent (incongruent) phase, and $N = n_1 + n_2$.

As illustrated in the right side of the D-equations, the *D*-measure can be considered as a scaled point-biserial correlation coefficient (r_{pb}), that is, a correlation between a dummy variable indicating the phase (e.g., '0' for congruent trials, and '1' for incongruent trials) and

a variable with response latencies. For those studies in which the same number of responses in both phases are analyzed, r_{pb} and *D* will yield similar psychometric results[2]. This allows us to infer specific hypotheses about the *D*-effect size measure based on studies discussing the effectiveness and properties of r_{pb}. Although the D-measure seems to be an improvement over C-measures (see Greenwald et al., 2003), the appropriateness of the measure given the properties of the data typically observed for an IAT (or generally, for response latencies in choice reaction time tasks) is scarcely discussed. For instance, one important problem is that, in general, response time distributions are typically skewed to the right, and therefore are sometimes modeled as exponentially modified Gaussian (ex-Gaussian) or Wald-distributions (Fazio, 1990; Van Zandt, 2002; Balota and Yap, 2011). Often, the data are therefore log-transformed or inversed (reciprocal transformation; Fazio, 1990; Ratcliff, 1993). However, even with a log transformation, these distributions are not always symmetrical in shape around some mode. Moreover, the inherent positive skewness makes it challenging to decide which observation can be treated as an *outlier*, thereby increasing researchers degrees of freedom (i.e., the freedom of the researcher to make specific decisions during the data-collection and data-analyses stages of research). Also, heteroscedasticity - a different degree of variation in both samples - is often observed for reaction time experiments (Grissom & Kim, 2001). Beside these distributional properties, the number of observations can vary between conditions. Although in most tasks, the number of trials in each condition will be balanced, analyzing only correct trials or excluding trials based on time boundaries might cause a different number of trials in each condition that are used to estimate the effect size score.

Several authors have discussed the appropriateness of effect size measures given the properties of data and (experimental) settings (e.g., Rozenthal, 1994; Grissom and Kim, 2001; McGrath and Meyer, 2006; Newcombe, 2006; Ruscio, 2008; Wilcox & Tian, 2011, Lakens, 2013). McGrath and Meyer (2006) have shown that r_{pb} is sensitive to a difference in the number of trials in each condition used to estimate the effect size. Different values were obtained by varying the number of trials sampled from the same distribution. Ruscio (2008)

[2] Greenwald et al. (2003) have named the D-measure "D" in order to refer to the well-known Cohen's *d*. However, in the literature on the properties of standardized effect size measures, Cohen's *d* and r_{pb} are considered as the prototypes of two different families of effect size measures with different properties (e.g., McGrath & Meyer, 2003). Because the IAT D-measure can be considered as a scaled r_{pb}, one could argue that the name "D-measure" should be replaced with "R-measure".

argued that r_{pb} is sensitive to deviations of normality distributions and homogeneity of variances. In addition, r_{pb} is very sensitive to outliers. For example, Grissom and Kim (2001) note that even '*slightly heavy tails in the distribution of the dependent variable scores can greatly affect* [r_{pb}]' (p139). Indeed, statistics such as the sample mean and sample variance are sensitive to outliers, or as formulated by Greenwald et al. (1998, p. 1467) "*they distort means and inflate variances*".

3.3 The Probabilistic Index

In the present paper, we introduce the Probabilistic Index (PI) as an effect size measure for existing and new scoring algorithms, and we propose to use Probabilistic Index Models (PIMs) for the estimation of the PI from IAT trials. PIMs are a new class of semiparametric regression models introduced by Thas, De Neve, Clement, and Ottoy (2012), The PI is defined as

$$P(Y \leq Y') \coloneqq P(Y < Y') + \frac{1}{2}P(Y = Y'),$$

where Y and Y' denote two independent responses associated with covariates X and X', respectively. The PI is a summary measure with a clear interpretation, namely, the probability that a randomly chosen response exceeds another randomly chosen response with regard to a certain response property (e.g., time needed to emit the response). More specifically, in an IAT it would reflect the probability that a randomly selected response on a congruent trial is faster than a randomly selected response on an incongruent trial. The PI is modeled by PIMs as a function of the covariates. PIMs are well suited for both ordinal and interval measurement scales. This allows us to consider response latencies as measured on an ordinal scale. In other words, it is sufficient to assume that a response latency of 800ms, for example, is slower than a response latency of 400ms, without making the additional assumption that the first response latency reflects a mental process that takes twice the amount of time to operate as the mental process that mediate the latter response. It is important to note that both C- and D-measures rely on this assumption.

Consider the PIM $P(Y \leq Y' \mid X, X') = \text{expit}[\beta_1(X' - X)]$,

where X is a dummy variable, with $X = 0$ for congruent trials and $X = 1$ for incongruent trials and expit = exp(x)/ [1 + exp(x)]. The equation can now be written as:

$$P(Y \leq Y' | X = 0, X' = 1) = \text{expit}(\beta_1).$$

From this expression, the interpretation of $\text{expit}(\beta_1)$ becomes clear: it is the probability that a randomly selected trial in the congruent block (B4) has a shorter response latency than a randomly selected trial in the incongruent block (B7). The result is an estimate of the IAT-scores that could vary between 0 and 1, with 0.5 the theoretical 'zero-point' . That is, a PI-score of 0.5 indicates an equal probability to respond faster or slower in the congruent block relative to the incongruent block. In contrast to the *D*-measure, the PI as defined for the two-sample case can be considered as the degree of overlap between the response latency distributions of the two blocks (in this case, the distributions of the congruent and the incongruent blocks) and is thus not defined as a (weighted) difference in mean reaction time (Newcombe, 2006).

The PI as defined above would be similar as estimating the effect based on the *A*-statistic described by Ruscio (2008). As illustrated by Ruscio (2008), this probability-based effect size measure can easily be estimated by scaling the Mann-Whitney U statistic or Wilcoxon test statistic W_m (see also Newcombe, 2006). However, by introducing a more general model, more complex designs can be explored within the same framework. The described *A*-statistic can thus be considered as an outcome of a specific model of the PIMs. In Appendix 1, we illustrate how the PI can be estimated using the statistical software R (R Core Team, 2014).

Based on the studies of Newcombe (2006) and Ruscio (2008), we may conclude that the PI is a better choice as an effect size measure for choice reaction time measures. For instance, the PI remains rather insensitive to differences in the number of trials observed in each condition. Perhaps more importantly, the PI remains (relatively) robust against the presence of outliers and still has a relevant interpretation in cases of non-normality and/or heteroscedasticity. In the present paper, we draw attention to the impact of outliers on both the D-measure and the PI-measure. Although we know that outliers have a large impact on the D-measure, less is known about *how* outliers might affect these measures. In order to evaluate and compare the usefulness of the PI-measure against the D-measure, we will propose two new

scoring-algorithms. It is important to note that the effect size measure and, more generally, the scoring algorithm is an integral part of any test. Hence, by changing the effect size measure or the scoring algorithm of a test, another test is obtained in the sense that different algorithms reflect different types of performance. Thus, it is not guaranteed that modified tests are still useful in other contexts. Therefore, the new scoring-algorithms and two D-scoring algorithms will then be applied on several datasets used to compare implicit attitude measures in three different domains.

3.4 Simulations of the impact of outliers on D-scores and PI-scores Methods

To examine the impact of outliers on the IAT-score, a Monte Carlo simulation was designed. In a first condition, we chose to generate data by sampling 40 *trials* two times from two normal [3] distributions $Y_1 \sim N(\mu_1 = 800, \sigma^2 = 100^2)$ and $Y_2 \sim N(\mu_2 = 800 + \delta_i, \sigma^2 = 100^2)$. These values assure that only positive numbers are obtained and that the outlier (defined as the mean score plus 6 times the standard deviation) will be a rare observation. The first sample (Y_1) represents the *congruent phase*, while the second sample represents the *incongruent phase* (Y_2). The effect-size is determined by δ_i, with δ varying from 5 to 250, with a step size of 5, resulting in 50 different effect-sizes. In the second condition, the value of the first trial of Y_1 (in essence, this is a random trial) is replaced by 1400 (μ_1 + 600). The value of μ_1 + 600 is chosen because this corresponds with the practice of treating error trials as determined by the popular D4-scoring algorithm (Greenwald, Nosek & Banaji, 2003). In a third condition, the value of the first trial of Y_2 is replaced by (μ_2 + 600), with $\mu_2 = 800 + \delta_i$. For the final condition both values of the first trials of Y_1 and Y_2 are replaced by μ_1 + 600 and μ_2 + 600 respectively. For each setting, 1000 Monte Carlo simulations runs are used and the effect size is estimated by both the D-measure and the PI-measure.

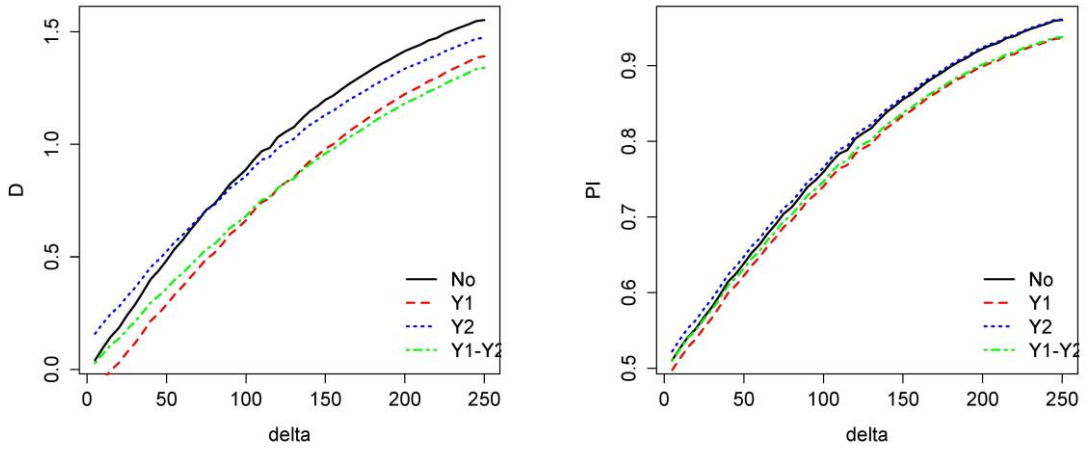
3.4.1 Results and Discussion

For each setting the mean score for the 1000 runs was estimated and are presented graphically in Figure 1. Table 2 summarizes these means for δ = 5, 125, 250. In this table, the

[3] This should be the most optimal setting for the D-measure.

D-scores were transformed by dividing the D-effect size score by 4 and adding 0.50. This transformation allows us to compare the effect of adding an outlier on a similar scale (note that the original D-scores are bounded by -2 and +2, while the PI is bounded by 0 and 1). Absolute deviations due to the presence of an outlier seem to be systematically larger for the D-measure compared to the PI-measure. For both measures, adding an outlier to the *congruent phase*, decreased the estimated scores. However, the obtained mean differences (i.e., the difference between the mean score of the scores obtained by the 1000 runs for a given δ in the outlier condition and the mean score for the 1000 runs for the same δ in the condition without an outlier) are almost two times higher for the D-effect size measure compared to the PI-measure. If an outlier was added to the *incongruent phase*, the impact differs for both measures. While for the PI, scores somewhat increase, the impact on the D-measure is a function of the effect size δ: for small effect sizes, higher scores are obtained, while for larger effect sizes, smaller scores are obtained compared to the scores of the non-outlier condition. This is remarkable, because the intended 'penalty' given by the D4-scoring algorithm might have a different impact depending on someone's true (i.e., the mean score obtained in the condition without outliers) effect size: a score will be overestimated when true effect sizes are small, and it will be underestimated when true effect sizes are large. Also, the absolute deviation between the estimated and the *true* effect size seems to be larger for the D-measure compared to the PI-measure. In the case of two outliers, one in each phase, the impact becomes even more clear for larger effect sizes. Again, for the PI-measure rather small deviations are observed. With respect to the D-measure, the outlier-effect tends to be largest (compared to the two previous conditions where only one outlier was defined) for large effect sizes.

In sum, our simulation demonstrated the impact of outliers on observed scores. As expected, the PI-measure is more robust against the presence of outliers relative to the D-measure. If an outlier was added to the incongruent condition, higher scores were obtained for true low effect sizes, while lower scores were obtained for true high effect sizes. Also, our results question the use of a fixed penalty as suggested by Greenwald et al. (2003) because the penalty might have an opposite effect depending on the true effect size score of participants.



A. B.

Figure 1. Estimated scores based on four different conditions. The "No" condition refers to the condition without outliers; "Y1" ("Y2") refers to the condition with an outlier in the congruent (incongruent) phase; "Y3" refers to the condition with outliers in both phases. D = effect size based on the D-measure (Left panel); PI = effect size based on the PI-measure (Right panel).

Table 2. Simulation results for exploring the impact of outliers, based on 1000 Monte Carlo runs.

	D*				**PI**			
δ	**con1**	**Δcon2**	**Δcon3**	**Δcon4**	**con1**	**Δcon2**	**Δcon3**	**Δcon4**
5	.510	-.033	.029	-.003	.511	-.013	.012	-.001
125	.765	-.056	-.012	-.056	.811	-.020	.005	-.015
250	.888	-.040	-.020	-.053	.960	-.024	.001	-.023

Note. D* = D/4+0.50; D = effect size based on the D-measure; PI = effect size based on the PI-measure. δ = effect size, defined as mean difference; con1 refers to the mean scores obtained in the sample without outliers; Δcon2 (Δcon3) refers to the differences between the mean score obtained in condition 2 (3) with an outlier in the congruent (incongruent) phase and the mean score obtained in con1; Δcon4 refers to the differences between the mean score obtained in the condition with outliers in both phases

Because the presence of outliers might be expected in empirical data and because the impact of outliers seems to be larger for IAT-scores based on a D-effect size measure compared to a PI-effect size measure, we expect that using the PI-measure instead of the D-measure can enhance the reliability of IAT measures.

3.5 Applying the measures on actual data

To evaluate the PI-scoring algorithm, we re-analyzed data from a study reported by Bar-Anan and Nosek (2014) in which they compared several implicit attitude measures across three different domains (i.e., the so-called "Attitudes 3.0" data set). We propose two new scoring-algorithms using the PI-effect size measure: the PI-scoring algorithm for IATs using built-in error penalties (registering the time until a correct response is given) and the PIe-scoring algorithm for IATs registering the time until a response (correct or incorrect) is given. As discussed in previous section, using the PI-effect size measure, a clear interpretation of the effect will be obtained. Also, we might expect an increase of the test reliability. It is important to note that the effect size measure and, more generally, the scoring algorithm is an integral part of any test. Hence, by changing the effect size measure or the scoring algorithm of a test, another test is obtained in the sense that different algorithms reflect different types of performance. Thus, it is not guaranteed that modified tests are still useful in other contexts. Therefore, it is important that we obtain at least similar results with respect to validity-related criteria.

3.5.1 The PI and PIe scoring algorithms

The D1-score as defined by Greenwald et al., (2003) is the average of the D-scores obtained in the practice and test blocks. All latencies are used for estimating the effect size, except those from trials with a response time larger than 10000ms. The D2-scoring algorithm uses the same strategy, except trials less than 400ms are removed. The D4-measure differs from the D1-measure only with respect to error treatment: before the effect size is estimated, latencies of error trials are replaced by the mean response latency of the trials belonging to the same block plus 600ms.

Similar to the D scoring algorithm, the PI and PIe scoring algorithm will remove latencies above than 10000ms. This arbitrary boundary was chosen by Greenwald et al., (2003) to avoid 'extravagant' latencies. Consider the PIM

$$P(Y \leq Y' | \boldsymbol{X}, \boldsymbol{X}') = \text{expit}[\beta_1(X'_1 - X_1)],$$

where X_1 is a dummy variable, with $X_1 = 0$ for a congruent trial (B3 or B4) and $X_1 = 1$ for an incongruent trial (B6 or B7). Let X_2 denote a second dummy variable, indicating the phase, with $X_2 = 0$ for practice trials and $X_2 = 1$for test trials. The PIM can be expressed as:

$$P(Y \leq Y' |\ X_1 = 0,\ X'_1 = 1,\ X_2 = X'_2) = \text{expit}[\beta_1(X'_1 - X_1)] = \text{expit}(\beta_1),$$

under the condition that both response latencies *Y* and *Y'* belong to the same phase.

The PIe-scoring algorithm is similar to the PI-scoring algorithm, with the difference that error trials are treated differently. By substituting the response latency of an error trial by a fixed penalty of 10000ms, response latencies of error trials are relocated to the end of the distribution (remember that observations larger than 10000ms are removed for all scoring algorithms before applying the penalties). In contrast to the D4 measure, the effect of a penalty will remain the same for each participant and for each IAT because it does not depend on the mean response latencies.

3.5.2 Reanalyzing the Attitude 3.0 data

The three IATs used in the study of Bar-Anan and Nosek (2014) are a Self-IAT (Self vs Others), a Race-IAT (White People vs Black People) and a Political-IAT (Democrats vs Republicans). For all three IATs the attribute labels were Good Words vs Bad Words. The structure of the three IATs is similar to the IAT outlined in Table 1. For each IAT, the order of the target and attribute combinations was randomized. For each IAT, we selected only participants who completed the IAT. For those participants who were tested twice, we only used the data of their first IAT. Participants with 10% response times less than 300ms were eliminated. This resulted in a final sample of $N = 2747$ for the Self-IAT, $N = 2894$ for the Race-IAT and $N = 2780$ for the Political IAT. Based on the raw data, we estimated for each participant their IAT-score by using the D2-scoring algorithm[4], the D4-scoring algorithm, the PI-scoring algorithm and the PIe-scoring algorithm.

[4] We selected this algorithm because the D2-scoring algorithm was chosen by the authors of the original study (Bar-Anan & Nosek, 2014). We also analyzed the data using the D1-scoring algorithm. Almost identical results were obtained for D1 and D2.

Reliability, approximated by internal consistency[5], was explored using the method described by Bar-Anan and Nosek (2014). These authors suggest that three parcels be created based on the consecutiveness of trials (for example, the first parcel consists of the first, fourth seventh, ... trials). We computed Cronbach's alpha from these parcels. In addition, we estimated the average interparcel correlation coefficient. Besides Cronbach's alpha, we also estimated the correlation between two test halves, using an odd/even split. Also, to avoid a distortion of the reliability measure by the effect of which category was paired with which attribute first, separate correlation coefficients were estimated for each order.

To explore the utility of the IAT, correlation coefficients were estimated with self-report measures[6]. Self-report measures were difference scores between explicit questions regarding the two attitude objects; 'feeling thermometers' ("how warm or cold do you feel toward the following group", 0 = coldest feelings, 5 = neutral, 10 = warmest feelings); self-reported preference ("which statement best describes your personal feelings toward group A and group B, 1 = strong preference of B over A, 7 = strong preference of A over B). For Race and Politics, participants were asked to rate content specific items (e.g., a picture of Bill Clinton; 0 = coldest feelings, 4 = neutral, 8 = warmest feelings). The final score was the difference between the mean item ratings for each attitude object. Also, for each attitude object, scores of a content-related questionnaire were applied: the Rosenberg Self-esteem (RSE) the Modern Racism Scale (MRS) and the Right-wing authoritarianism (RWA). Finally, participants were asked which candidate they had voted for (Voted) and which candidate they would vote for (Will Vote). In line with Bar-Anan and Nosek (2014), average correlations were calculated for each topic and scoring algorithm.

To explore the predictive value of the Political-IAT, we examined the relation between the Political IAT-scores given the D2, D4, PI and PIe-scoring algorithms and self-reported measures of voting behavior (Voted and Will Vote). Several strategies are possible to examine the predictive value of a test. The simplest way to evaluate measures is to explore the *accuracy*

[5] Internal consistency might refer to different concepts (e.g., homogeneity or unidimensionality; see Sijtsma , 2009; Clark and Watson, 1995; and Yang and Green, 2011, for a discussion). Here, internal consistency refers to the interrelatedness of different a priori defined parcels.

[6] The self-report measures used in this study can be inspected via the following link: https://dw2.psyc.virginia.edu/implicit/user/yba/mtmmr/mtmmselfreport.htm

of the test. With accuracy, we refer to the (average) number of correct predictions given a well-defined *decision threshold*. Suppose zero is the theoretical decision bound for the Political IAT using a D-scoring algorithm, then cases with positive scores can be classified as favoring democrats, while cases with negative scores can be classified as favoring republicans. Because the 'true' voting behavior is known (assuming that participants did answer the question truthfully), we easily can calculate the accuracy of the test. However, exploring accuracy - as defined above - might be misleading because this number strongly depends on the number of actual democrats or actual republicans in the sample or population (see Metz, 1978). Therefore, accuracy will be explored via the concepts *sensitivity* and *specificity*. In this study, a positive score theoretically indicates a preference for democrats over republicans. This means that in this case, sensitivity refers to the percentage of 'true' democrats decisions, given the total number of actual democrats in our sample. Specificity refers to the percentage of 'true' republicans decisions, given the number of actual republicans in our sample. Scoring algorithms with higher sensitivity and specificity are preferred. Because IAT-scores could yield more certainty to vote for either democrats or republicans in comparison to scores closer to the theoretical zero-point (note that this zero point equals 0.5 for the PI), we will also calculate the Brier score[7] (Brier, 1950, Wilks, 2011). Scoring algorithms with smaller Brier scores are preferred. Finally, because some authors question the zero-point of an IAT (Blanton & Jaccard, 2006), we graphically explored the effect of varying decision thresholds. Therefore, Receiver Operating Characteristics (ROC) curves are plotted to examine the predictive value (Metz, 1978). ROC curves plot the sensitivity against 1-specificity for varying decision thresholds. Scoring algorithms with a ROC curve passing through the left-upper corner (i.e., only true positive and no false positives) of the ROC space are preferred (Metz, 1978).

[7] The Brier score is defined as $BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2$, with p_i the probability that an event will occur and $o_i = 1$ if the event occurred (here, voting for democrats) and $o_i = 0$ if the event not occurred (here, voting for republicans) and N the number of participants. For example, suppose participant x has a PI-score of 0.9. We predict that this participant would vote for the Democrats. If we are correct, the Brier score is $(0.9 - 1)^2 = 0.01$. On the other hand, if this participant voted for the Republicans, the prediction-error is $(0.9 - 0)^2 = 0.81$.

3.5.3 Results

Internal consistency. Table 3 presents the estimated Cronbach's alphas, interparcel correlation coefficients for the Self IAT, Race IAT and Political IAT and the estimated split-half correlation coefficients for their two specific versions. Across IATs, internal consistency measures were systematically higher for the PI-scoring algorithms. Even in those cases where internal consistency was already high, an increase was obtained by using the PI-effect size measure. The PIe-scoring algorithm, which includes an error-correction resulted in only a small decrease relative to the PI-scoring algorithm. If we contrasted the D4-measure with the PIe-measure, the latter seems to be more internal consistent.

Table 3. Summary of the internal consistency results.

IAT	D2	D4	PI	PIe
Self				
alpha	.83	.79	.89	.87
IR	.62	.56	.72	.70
SH self-first	.62	.58	.74	.72
SH other-first	.62	.56	.72	.70
Race				
alpha	.86	.84	.91	.90
IR	.68	.65	.77	.75
SH white-first	.68	.65	.75	.73
SH black-first	.67	.62	.75	.74
Political				
alpha	.93	.91	.95	.94
IR	.82	.78	.86	.85
SH dem-first	.79	.75	.83	.82
SH rep-first	.82	.77	.86	.84

Note. alpha = coefficient alpha based on three parcel;. IR = Interparcel correlation coefficient; SH = split-half correlation coefficient.

Correlations with explicit measures. Table 4 summaries the Pearson correlation coefficients for each direct measure with the related IAT-scores estimated by the D2, D4, PI and PIe-scoring algorithm. No substantial differences were observed between the different scoring algorithms. The average correlation coefficients for the Self-esteem and Political related self-report measures with the IAT-scores did not vary among different scoring

algorithms. For the Race IAT, slightly better results were obtained if IAT-scores were estimated using a D-effect size measure.

Table 4. Pearson correlation coefficients between IATs and direct measures

IAT	D2	D4	PI	PIe
Self				
Therm (n=598)	.13	.12	.14	.14
Pref (n=598)	.10	.10	.09	.09
RSE (n=535)	.18	.18	.20	.20
Mean	.14	.13	.14	.14
Race				
Therm (n=593)	.33	.32	.31	.31
Pref (n=580)	.33	.32	.30	.30
Items (n=632)	.21	.22	.19	.19
MRS (n=621)	.29	.28	.27	.27
Mean	.29	.29	.27	.27
Political				
Therm (n=548)	.61	.60	.61	.61
Pref (n=544)	.65	.64	.64	.64
Items (n=423)	.68	.69	.68	.68
RWA (n=534)	.46	.47	.46	.46
Voted (n=286)	.68	.68	.70	.69
Will Vote (n=516)	.52	.53	.53	.53
Mean	.60	.60	.60	.60

Prediction. Inspection of Table 5 shows that the PI-scoring algorithms slightly increase the predictive value of the Political-IAT of past and future voting behavior: larger values for sensitivity and specificity are obtained for scores obtained by the PI and PIe compared to D2 and D4-scores respectively. Furthermore, the Brier score for past voting and future voting were smaller when PI-effect size measures were used.

Figure 2 shows the ROC spaces for voting behavior based on the Political-IAT using different effect size measures. We contrasted the D2-scoring algorithm against the PI-scoring algorithm (top panels of Figure 2) and the D4-scoring algorithm with the PIe-scoring algorithm (bottom panels). When good prediction is obtained by all measures (left panels of Figure 2), only a slight difference is observed between the curves based on a PI or D-effect size measure. The curves of the PI are attracted to the right and top of the ROC space to a slightly larger extent. A similar, but more clear trend can be observed on the right panels of Figure 2. The ROC curves indicate that better tests are obtained when a PI-scoring algorithm is used instead of a D-measure.

Table 5. Predictive strength of Political IATs.

IAT	D2	D4	PI	PIe
Voted				
Sensitivity	.91	.93	.93	.93
Specificity	.69	.72	.72	.74
Brier score	.16	.15	.13	.13
Will Vote				
Sensitivity	.83	.84	.85	.85
Specificity	.71	.71	.75	.74
Brier score	.17	.17	.15	.15

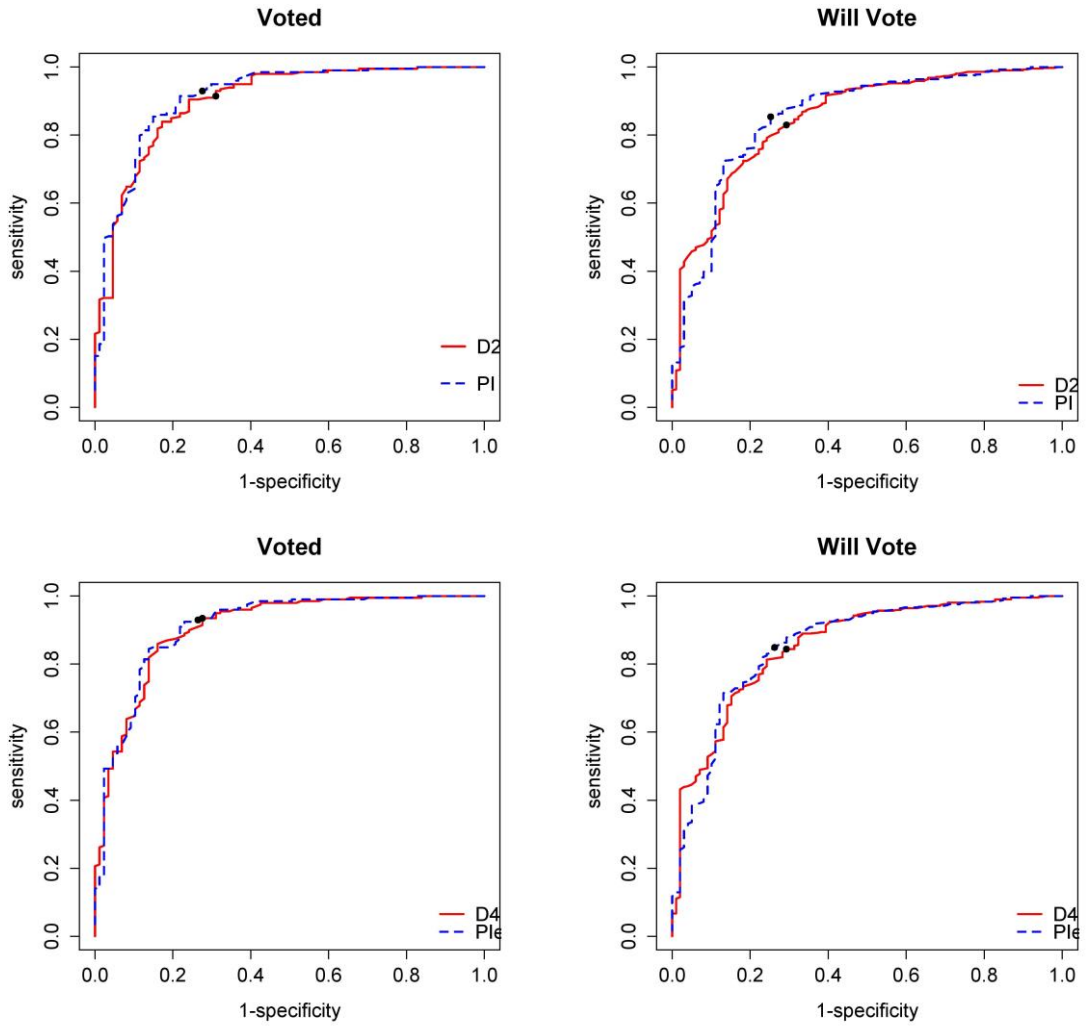


Figure 2. ROC space and curves based on the D2 and PI scoring algorithm (Upper panels) and D4 and Pie (Lower panels). A Black dot indicates the position on the curve when the theoretical zero-point is used as cut-off score.

3.6. General Discussion

We introduced the PI effect size measure as a theoretically and practically sound alternative to existing measures used to estimate IAT-scores. Because the IAT is used increasingly often in applied contexts (e.g., see Greenwald, Poehlman, Uhlmann, & Banaji, 2009 for a review), it is important to optimize the way in which IAT-scores are approximated. Inspired by the properties of different standardized effect size measures and recommendations concerning the applicability of these measures (given the distributions of the observed data; Ruscio, 2008), we examined how D- and PI-measures are affected by the presence of outliers. Furthermore, we re-analyzed the data from three different IAT studies to compare performances of D- and PI-measures on different criteria: reliability, correlation with explicit measures, and predictive validity. Our results showed that the PI has several important theoretical and practical advantages. Furthermore, it is easily calculated and interpreted.

While the D-measure was sensitive to the presence of outliers, the PI-measure was relatively robust to outliers. It must be noted that this is a very valuable characteristic, as outliers are often present in, for instance, clinical data (e.g., Wiers et al., 2007) and in data collected online (Greenwald et al., 2003). Also, we observed a systematic improvement in the estimated reliability when IAT data from three different attitude domains were re-analyzed using the PI scoring algorithm. Similar correlation coefficients where observed between the IATs and their related explicit measures. Only for the Race-IAT did scores based on the PI-measure correlate less strongly with direct measures compared to the coefficients obtained from the D-measures. Furthermore, using the PI measures improved the predictive value of a Political-IAT: both the sensitivity and the specificity of the IAT increased, while the Brier score decreased when the PI measure was used. In addition, ROC curves based on the PI-measure were slightly superior than ROC curves based on D-scores. Together these results suggest that there are important advantages to using the PI over the D-algorithm when estimating IAT-scores.

Although the introduction of the D-effect size measure by Greenwald et al. (2003) significantly decreased the impact of general response speed, it came with a cost: compared to previous effect size measures, less measurement precision was obtained, reflected in lower

internal consistency or reliability estimates (Greenwald et al., 2003). Given the possibility of extreme response times in IAT data and right skewed distributions, we argued that by using an effect size measure such as the PI that is robust to the presence of outliers, it should be possible to increase measurement precision for a same set of observations. Indeed, applying the PI-effect size measure did increase the reliability for several IATs.

Besides reliability, other important dimensions of psychometric evaluation could be affected by changing the scoring algorithm. To explore this, we examined the utility and the predictive value of D and PI indices. The results of the correlation studies were similar for the D-related measures and the PI-related measures, suggesting that, according to this criterion, both measures are equal. It is important to note, however, that correlations between implicit measures and explicit measures reveal the degree of relatedness (as measures of convergent validity), but at the same time reveal how well implicit measures can be distinct from explicit measures (as measures of discriminant validity). If we strive for high convergent validity, high correlations between implicit and explicit measures are desirable, but if we strive for high discriminant validity, low correlations are desirable. Choosing one of these types of validity over another as a criterion to decide which measure should be preferred seems to be a rather subjective choice. Our results must reassure the 'convergent' believers that by using a different effect size measure, the degree of relatedness between the implicit and explicit measures is at least the same for the Political-IAT and Self-IAT. On the other hand, 'discriminant' believers could be satisfied that the correlation between the Race-IAT scores and explicit scores is somewhat smaller when using the PI measures. More importantly, in situation where prediction and evidence is important, best results were obtained for the IAT using the PI-effect size measure.

Apart from the psychometric merits of using a PI-effect size measure, the PI has some other favorable properties. Because the PI treats response times as measured on an ordinal scale, it allows an elegant way to handle error-trials for IATs without a build-in error penalty procedure. The popular D4-measure changes the response time of an error trial by replacing the score with the block mean response time plus 600 ms, causing different and undesired effects within and between participants (e.g., the penalty might increase or decrease the total score, depending on the size of the *true* score). By giving a penalty of 10000 ms, error trials

are relocated at the end of the distribution. For each participant, each IAT and in each setting, the effect will be similar: an error trial will be treated as the slowest trial possible. Moreover, our results suggest that scores based on the PIe-scoring algorithm approximate the scores as obtained by the PI-scoring algorithm, indicating that this procedure mimics the build-in error penalty procedure (i.e., it is most likely that reaction times of corrected responses will be located at the end of the distribution).

Most importantly, however, the PI-score has the advantage that it has a clear interpretation. The PI is useful in situations where non-experts want an answer on a simple ordinal question (for instance, "Was I faster in responding on congruent trials than on incongruent trials"? "Indeed, the estimated probability that a randomly selected trial in a congruent phase has a smaller response latency than a randomly selected trial in a incongruent phase equals 80%"). This property should lead to an increase of the face validity of the IAT.

To conclude, in the present paper we introduced the PI as a new effect size measure for calculating IAT-scores that is easy to interpret. We proposed two new scoring algorithms, the PI and the PIe, based on the existing and popular D-scoring algorithms. By enhancing the process of data evaluation for an IAT, the measurement precision of the IAT can be improved – without decreasing the utility of the IAT.

References

Arcuri, L., Castelli, L., Galdi, S., Zogmaister, C., & Amadori, A. (2008). Predicting the vote: Implicit attitudes as predictors of the future behavior of the decided and undecided voters. *Political Psychology, 29,* 369-387.

Balota, D.A., and Yap, M.J. (2003). Moving beyond the mean in studies of mental chronometry: the power of response time distributional analyses. *Current Directions in Psychological Science, 20,* 160-166.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. Zeitschrift für Experimentelle Psychologie, 48, 145-160

Bar-Anan, Y. & Nosek, B.A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods, 46,* 668-688.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61,* 27-41.

Brier, G.W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review, 78,* 1-3.

Cunningham, W.A., Preacher, K.J., & Banaji, M.R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science, 12,* 163-170.

De Neve, J. (2013). *Probabilistic index models*. Unpublished doctoral dissertation. Ghent University. Faculty of Sciences, Ghent, Belgium.

Faust, M.E., Balota, D.A., Spieler, D.H., and Ferraro, F.R. (1999). Individual differences in information processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin, 125,* 777-799.

Fazio, R.H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick & M.S. Clark (Eds.), *Research methods in personality and social psychology* (pp. 74-97), Newburry Park, CA:Sage.

Friese, M., Bluemke, M., & Wanke, M. (2007). Predicting voter behavior with implicit attitude measures: The 2002 German parliamentary election. *Experimental Psychology, 54,* 247–255.

Greenwald, A.G., McGhee, D.E., and Schwartz, J.K.L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74,* 1464-1480.

Greenwald, A.G., Nosek, B.A., and Banaji, M.R. (2003). Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm. *Journal of Personality and Social Psychology*, *85*, 197-216.

Greenwald, A.G., Poehlman, T. A., Uhlmann, E. L., and Banaji, M.R. (2009). Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity. *Journal of Personality and Social Psychology*, *85*, 197-216.

Grissom, R.J., & Kim, J.J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods*, *6*, 135-146.

Kelly, K., & Preacher, K.J. (2012). On effect Size. *Psychological Methods*, *17*, 137-152.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*:863 doi: 10.3389/fpsyg.2013.00863.

Levy, B.R., & Banaji, M.R. (2002). Implicit ageism. In T.D. Nelson (Ed.), Ageism: Stereotyping and prejudice agains older persons. Cambridge, MA: The MIT press.

Maison, D., Greenwald, A. G., & Bruin, R.H. (2004). Predictive validity of the Implicit Association Test in studies of brands, consumer attitudes and behavior. *Journal of Consumer Psychology*, *14*, 405–415.

McGrath, R.E., & Meyer, R.E. (2003). When effect sizes disagree: The case of r and d. *Psychological Methods*, 11, 386-401.

Metz, C.E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8, 283-298.

Newcombe. R.G. (2006). A deficiency of the odds ratio as a measure of effect size. *Statistics in Medicine*, *25*, 4235-4240.

Olson, M.A., & Fazio, R.H. (2004). Reducing the influence of extrapersonal associations on the Implicit Association Test: personalizing the IAT. *Journal of Personality and Social Psychology*, *86*, 653-667.

R Development Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/ (ISBN 3-900051-07-0)

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510-532.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, 13, 19-30.

Rudman, L.A., Greenwald, A.G., & McGhee, D.E. (2001). Implicit self-concept and evaluative implicit gender stereotypes: Self and ingroup share desirable traits. *Personality and Social Psychology Bulletin, 27*, 1164-1178.

Sijtsma, (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*, *74*, 107-120.

Sriram, N., Greenwald, A.G. (2009). The Brief Implicit Association Test. *Experimental Psychology*, *56*, 283-294.

Thas, O., De Neve, J., Clement, L., & Ottoy, J.P. (2012). Probabilistic index models (with Discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74*, 623-671.

Van Zandt, T. (2002). Analysis of response time distributions. In J. T. Wixted (Vol. Ed.) & H. Pashler (Series Ed.) Stevens' Handbook of Experimental Psychology (3rd Ed), *Methodology in Experimental Psychology* (Vol.4, pp. 461-516). Springer, New York, USA.

Wiers, R.W., Teachman, B.A., & De Houwer, J. (2007). Implicit cognitive processes in psychopathology: An introduction. *Journal of Behavior Therapy and Experimental Psychiatry, 38*, 95-104.

Wilcox, R.R., & Tian, T.S. (2011). Measuring effect size: a robust heteroscedastic approach for two or more groups. *Journal of Applied Statistics, 38*, 1359-1368.

Wilks, D.S. (2011). *Statistical methods in the atmospheric sciences* (3rd edition). Amsterdam: Academic Press.

Yang, Y., & Green, S.B. (2011). Coefficient Alpha: A Reliability Coefficient for the 21st Century? *Journal of Psychoeducational Assessment, 29*, 377-392.

Chapter 4

The PI_{IRAP}: An alternative scoring algorithm for the IRAP

Maarten De Schryver, Ian Hussey, Jan De Neve, Aoife Cartwright,
and Dermot Barnes-Holmes

Abstract

The Implicit Relational Assessment Procedure (IRAP) has been used to assess the probability of arbitrarily applicable relational responding or as an indirect measure of implicit attitudes. To date, IRAP effects have commonly been quantified using the D_{IRAP} scoring algorithm, which was derived from Greenwald, Nosek and Banaji's (2003) *D* effect size measure. In the article, we highlight the difference between an effect size measure and a scoring algorithm, discuss the drawbacks associated with *D*, and propose an alternative: a probabilistic, semiparametric measure referred to as the Probabilistic Index (Thas, De Neve, Clement, & Ottoy, 2012). Using a relatively large IRAP dataset, we demonstrate how the PI is more robust to the influence of outliers and skew (which are typical of reaction time data). Finally, we conclude that PI models, in addition to producing point estimate scores, can also provide confidence intervals, significance tests, and afford the possibility to include covariates, all of which may aid single subject design studies.

4.1 Introduction

The first purpose of the current paper is to consider the relative benefits of effect size measures when scoring data from the Implicit Relational Assessment Procedure (IRAP: Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010), a reaction time task that is frequently employed within research related to Relational Frame Theory (RFT: Hayes, Barnes-Holmes, & Roche, 2001). The second purpose is to propose a probabilistic, semiparametric measure referred to as the Probabilistic Index (Thas, De Neve, Clement, & Ottoy2012) for use with the IRAP, which appears to provide some advantages over the currently most widely used measure (i.e., the *D*-IRAP score). Before proceeding, however, it seems important to provide a brief overview of the IRAP, as an RFT-based methodology, to contextualize the current work.

4.1.1 Historical and conceptual background to the IRAP

Much of the early research in RFT consisted of demonstration studies to test the theory's basic assumptions and core ideas. One of the defining features of this research was a dichotomous approach to arbitrarily applicable relational responding (AARRing), which is a central idea within the account of human language and cognition provided by RFT (see Hughes & Barnes-Holmes, 2016, for an accessible overview). That is, laboratory studies in RFT often focused on showing that particular patterns of AARRing were either present or absent. Within a few years of the publication of the 2001 RFT book (Hayes et al., 2001), however, the need to develop procedures that could, in principle, provide a measure of AARRing that was non-dichotomous became increasingly apparent. The initial response to this need was the development of what came to be known as the IRAP. Specifically, the IRAP was a response to the question, "How can we capture relational frames in flight", which essentially is a question about the relative strength of AARRing in the natural environment.

In developing the IRAP, two separate methodologies were combined. The first of these was an RFT-based procedure for training and testing multiple stimulus relations, the Relational Evaluation Procedure (REP: Cullinan, Barnes, & Smeets, 1998) and the second was the Implicit Association Test (IAT: Greenwald, McGhee, & Schwartz, 1998). The REP presents

participants with two stimuli and requires them to provide a relational response (e.g., "same" or "different"). The IAT was developed by social-cognition researchers as a method for measuring what are frequently conceptualized as associative strengths in memory by comparing the relative speed of categorization of stimuli. The IRAP combined features from these two tasks by requiring participants to provide one relational response in some blocks (e,g., "same") and another in other blocks (e.g., "different"), and comparing the relative speed of relational responding between block types. The IRAP was therefore conceptualized as a procedure for measuring the relative strength of AARRing in a non-dichotomous manner (see Barnes-Holmes, Hayden, Barnes-Holmes, & Stewart, 2008).

It has been argued that due to its close connection to the IAT research with the IRAP quickly became dominated by studies focused on so-called implicit attitudes and implicit cognition more generally (Barnes-Holmes, Barnes-Holmes, Hussey, & Luciano, 2016). On the one hand, this strategy was very useful because it provided a means by which to assess the validity of the IRAP as a measure of natural verbal relations (see Vahey, Nicholson, & Barnes-Holmes, 2015). On the other hand, it also served as a distraction from a focus on RFT and AARRing *per se* (Barnes-Holmes, Barnes-Holmes, Barnes-Holmes, Luciano, & McEnteggart, 2017). Furthermore, the historical connection between the IRAP and IAT was instrumental in developing a version of the IAT D_1 score, which is used to analyze the response latency data from the IAT. The IRAP version, the D_{IRAP} algorithm, is described later in the current article, but as was pointed out by Barnes-Holmes et al. (2010) the D_{IRAP} algorithm should not be seen as prescriptive or necessarily the "best way" to analyze IRAP data." (p.533). Consistent with this view, and the ongoing development of the IRAP as an RFT-based method for analyzing human language and cognition, the current article presents another algorithm for analyzing IRAP data that appears to offer a number of advantages over the D_{IRAP} algorithm.

4.1.2 A brief description of the IRAP

The IRAP is a computer-based task on which an individual responds to a series of trials, each of which usually presents pairs of stimuli on screen (although see Kavanagh, Hussey, McEnteggart, Barnes-Holmes, & Barnes-Holmes, 2016, for an alternative format using natural language statements). To illustrate, we use an IRAP that was designed to assess gender

stereotypes. (see Cartwright, Hussey, Roche, Dunne, & Murphy, 2017 for similar IRAPs and discussion of the topic). Subsequently, data collected using this IRAP will then be presented. On each trial a label stimulus appears at the top of the screen, such as either "Men are" or "Women are". Target stimuli appear in the middle of the screen, such as stereotypically masculine traits (witty, competitive, decisive, and charismatic) or feminine traits (nurturing, gentle, affectionate, and sensitive). Two response options are also provided on each trial, such as "true" and "false". The IRAP operates by requiring opposite patterns of responding across successive blocks of trials. For example, "men are-masculine" trials would require participants to respond with "true" on one block and "false" on the next block. If the correct response is emitted the task simply continues to the next trial, but if the incorrect response is emitted a red X appears on screen and the next trial is not presented until the correct response option is provided. The IRAP thus involves presenting four trial types within each block and participants are required to emit opposing patterns of responding across successive blocks of trials. The four trial-types for the example of the IRAP described above may be summarized as: men-masculine, men-feminine, women-masculine, and women-feminine. For half of the blocks of trials, participants would be required to respond as if men are masculine and women are feminine (consistent trials [1] ; i.e., men-masculine/true; men-feminine/false; women-masculine/false; women-masculine/true), and for the remaining blocks to respond as if men are feminine and women are masculine (inconsistent trials; i.e., men-masculine/false; men-feminine/true; women-masculine/true; women-masculine/false). Finally, it is worth noting that the IRAP typically involves allowing the participant to complete a number of pairs of consistent and inconsistent blocks until they reach mastery criteria (e.g., for each block in a pair, median latency < 2000 ms and accuracy > 80%), followed by a static number of test

[1] We employ the terms "consistent" and "inconsistent" here based on their usage in the literature. However, we recognize that these terms are potentially confusing. Given that Barnes-Holmes et al. (2010) stated that the faster/more probable response is by definition consistent with an individual's learning history, one could argue that a block of trials should only be designated as "consistent" for that individual after the fact (i.e., as an outcome) based on which block was faster, rather than a priori. As such, consistent and inconsistent are typically used in two ways: to note the (in)congruence between an individual's learning history and which block produces faster RTs, and as a label to differentiate the two types of blocks within the IRAP as the researcher sees them (likely influenced by social-normative expectations). These meanings will not always overlap, leading to occasional confusion. Here, we employ the latter sense of the words. For this reason, some researchers have referred to the blocks using arbitrary designations such as "A" and "B" (e.g., Hussey, Barnes-Holmes, & Booth, 2016) or as "pro" versus "anti" the domain of interest targeted by the IRAP (as in pro- and anti-spider; e.g., Nicholson & Barnes-Holmes, 2012)

blocks pairs (usually 3) from which data are analyzed. This was the case for the gender IRAP dataset used in the current paper (for paper length discussions of the task see Barnes-Holmes et al., 2010; Hussey, Thompson, McEnteggart, Barnes-Holmes & Barnes-Holmes, 2015).

Broadly speaking, the IRAP is usually scored by subtracting the mean response latency for one pattern of responding from the mean response latency of the opposite pattern of responding; the difference score is typically normalized (i.e., the D_{IRAP} algorithm). The difference score thus reflects a response bias in one direction or the other, such as responding "True" more quickly than "False" across blocks of trials when presented with the men-masculine trial-type. Specific response biases are usually predicted based on the behavioral histories of the participants. In the case of the current example, participants who report strongly gender-stereotypical biases would, for example, be expected to produce a larger men-masculine response bias. To put it simply, the original basic hypothesis behind the IRAP is that, all things being equal, *average or mean* response latencies should be shorter across blocks of trials that are consistent with a participant's behavioral history relative to those blocks of trials that require responses that are inconsistent with that history (see Barnes-Holmes, Finn, McEnteggart, & Barnes-Holmes, in press, for a recent and more sophisticated approach to explaining IRAP effects). In what follows, we will explain why the focus on the mean or average latency, which are used to calculate D_{IRAP} scores, may be problematic and outline an another analytic method for the IRAP. We will start by considering the general issue of scoring reaction time measures, and making an important distinction between the concept of a "scoring algorithm" and an "effect size measure".

4.1.3 Scoring reaction time measures

Since the introduction of the IRAP, IRAP scores have most frequently been calculated using the D_{IRAP} scoring algorithm. A scoring algorithm typically contains a set of consecutive steps that a researcher follows in order to obtain a final score, or scores, for each participant. For instance, a scoring algorithm might specify which trials should be taken into account, how to treat errors, how to treat response latencies that are deemed to be excessively short or long, and how to calculate the final score(s). Calculating the final score(s) usually involves adopting a particular effect size measure, which may be defined as the mathematical formula used to

calculate the quantity reflecting the magnitude of the difference in performance between conditions (e.g., between blocks of consistent and inconsistent trials). A specific scoring algorithm thus includes the type of effect size measure that should be used to obtain the score, or scores, in addition to other steps such as data exclusions. For instance, the D_{IRAP} scoring algorithm requires that all RTs > 10000 ms are discarded and then, for each trial-type for each pair of consistent and inconsistent blocks, a *D* effect size measure should be calculated. The means of the obtained *D* scores (calculated across the block pairs) serves as the final D_{IRAP} score for each trial type. Other D_{IRAP} scoring algorithms could propose similar steps, but include more stringent exclusion criteria than simply removing RTs > 10000ms, such as removing entire data sets for participants who failed to maintain a mean response latency < 2000ms on one or more of the four trial-types. Of course, other scoring algorithms could employ a different effect size measure. In the current article we propose one such measure: a semi-parametric probabilistic index.

Before proceeding, it should be noted that the decision to employ a particular effect size measure may be based in part on the stringency of the exclusion criteria. For example, if particularly stringent exclusion criteria are adopted for removing relatively long latencies, using an effect size measure that aims to reduce the impact of such latencies may be of little benefit. The basis for deciding how stringent the exclusion criteria should be is a highly complex issue, a discussion of which is beyond the scope of the current article. For present purposes, we will focus on a situation in which the exclusion criterion for response latencies is relatively relaxed (i.e., > 10000ms). As we shall see, when such a relaxed criterion is adopted, relatively long response latencies (e.g., lying somewhere between the mean response latency plus 2.5 standard deviations and 10000ms) may introduce unwanted "noise" into the dataset and thus it may be wise to adopt an effect size measure that will reduce the effects of such "noise"[2]. In what follows, we will begin by providing a concise overview of different types of effect size measures and discuss how appropriate these are to answer the main question: "is an individual faster (or slower) to respond on consistent trials compared to inconsistent trials?"

[2] We intentionally use quotes to indicate that extreme values in the context of some research questions may not be considered "noise" but constitute data points that have important theoretical significance, and should not, therefore, be removed from the dataset.

4.1.4 Non-standardized effect size measures

Non-standardized effect size measures summarize differences between distributions in the same unit as the unit of measurement. For instance, we can estimate the mean response time (RT, in milliseconds, ms) across consistent trials and the mean RT (also in ms) across inconsistent trials. The difference between these two means expresses the differences in ms between the "typical" RT for each block of consistent and inconsistent trials. Because RT distributions are typically right-skewed, other measures of central tendency are sometimes used, such as the median. Another way to deal with skewness is by transforming RTs using a log-, square root- and/or reciprocal transformation (e.g., the *C*-measures originally used for the IAT: Greenwald et al., 1998).

Non-standardized effect size measures have the advantage that they are easy to interpret, particularly when the unit of measurement (e.g., such as the RT rather than a difference score) is important. A serious limitation, however, is that effects tend to correlate with general responding speed (GRS); that is, participants responding slower during a task show typically larger effects compared to participants with faster responses (Fazio, 1990; Faust, Balota, Spieler, & Ferraro, 1999; Greenwald et al., 2003). Relatedly, O'Toole and Barnes-Holmes (2009) reported that raw latency and difference scores from the IRAP correlated with intelligence scores. This makes it difficult, or even impossible, to make meaningful comparisons of non-standardized effect sizes among participants, and even among different experiments.

4.1.5 Standardized effect size measures

Standardized effect size measures may be seen as "canceling out" the unit of measurement. Perhaps the most well-known of these is the standardized mean difference, or Cohen's *d*. Because the difference between the means is divided by the pooled standard

deviation, Cohen's *d* can be interpreted as the difference relative to the variability on RTs between conditions[3].

In their original article, Greenwald, et al. (2003, p. 201) explicitly draw a link between the *D* effect size measure and Cohen's *d*, both in terms of their calculation and interpretation (Cohen, 1988). However, it is important to note that *D* is, in actuality, mathematically more comparable to a different standardized effect size measure: the point-biserial correlation (r_{pb}) coefficient (see Ruscio, 2008). As will be discussed below, this categorization may be important when considering the disadvantages associated with different classes of effect size measures. Point-biserial correlations are expressed as the correlation between the RTs of both condition and a dummy variable indicting to which condition the RT belongs (0 for consistent responses; 1 for inconsistent responses). If the number of trials are equal in both conditions ($n_1 = n_2$), the *D* can be considered as a scaled point-biserial correlation coefficient*:* $r_{pb} = D \times \sqrt{\frac{n_1 n_2}{N^2 - N}}$, with $N = n_1 + n_2$. In contrast to Cohen's *d*, the *D* effect size measure is obtained by dividing the difference of the means by the standard deviation of the pooled sample of RTs (i.e., the standard deviation of RTs independent of condition; in the case of the IRAP, this means calculating single standard deviations across pairs of blocks of consistent and inconsistent trials). Broadly speaking, as measures of effect size, both *d* and r_{pb} can both be interpreted as a signal to noise ratio (i.e., in the differences in mean RTs between consistent and inconsistent blocks proportionate to the variance in those RTs).

Both *d* and r_{pb} effect size measures are popular, and both appear to reduce the unwanted correlation between effect sizes and GRS. However, they also have disadvantages in common (see Ruscio, 2008). First, r_{pb} seems to be sensitive to base rates. That is, when the number of trials substantially differs among conditions, r_{pb} will decrease in value if the difference in number of trials increases. In those cases where heterogeneity exists, the same observation is made for Cohen's *d*. Second, both measures are sensitive to violations of parametric assumptions such as normality and heterogeneity of variances. For example, Cliff (1993) argued that it is not guaranteed that if the mean of a first distribution is larger than the mean of

[3] Note, that some researchers divide the difference between the means by the standard deviation of a reference group. This standardized effect size measure is known as Glass d (d_G) and should not be confused by Cohen's *d*.

a second distribution, the majority of scores of the first distribution will be larger compared to the second one. For instance, this could be the case when the mode of a right heavy-tailed distribution is smaller compared to the mode of another (see also McGraw & Wong, 1992). Nonlinear data transformations, such as log-, square root-, and/or reciprocal transformations, have often been suggested as a way to correct for the non-normal distribution of RT data. Third, both Cohen's *d* and r_{pb} are sensitive to nonlinear transformations of the data. As such, different values might be obtained when one of the aforementioned transformations was used prior to the calculation of effect sizes. Fourth, both effect size measures are very sensitive to the presence of outliers. Finally, r_{pb} has the disadvantage that it is difficult to interpret, especially for non-experts.

At this point, let us focus on the potential impact of outliers and how we might deal with them in a set of IRAP data (as noted earlier, RT distributions are typically skewed to the right and this makes it difficult to decide which observations may be considered outliers). The data were taken from an unpublished IRAP study that was designed to examine gender stereotyping. In our sample (N = 188), we observed that 85% (81%) of the distributions of the consistent (inconsistent) trials show a skew to the right (here, we defined skewness when the Pearson's moment coefficient of skewness > 1.00). To test if the variances of the consistent block significantly (with alpha set to .05) differ from the variance of the inconsistent block, we performed F-ratio tests. Our results suggest that in 45% of the cases, unequal variances were observed (i.e., heterogeneity of variances). Next, we counted the number of outliers for each participant and for each type of block (i.e., consistent versus inconsistent). Here, any observation that differs at least 2.5 standard deviations from the mean (within participant, within block) is considered an outlier[4]. Our results show that at least one trial out of 24 trials can be identified as an outlier in 85% (79%) of the participants for the consistent (inconsistent) trials. In total, 3.8% of the trials were considered as outliers[5].

[4] We fully recognize that 2.5 SD is an arbitrary choice to define outliers (as is excluding RTs > 10000 ms in the D_{IRAP} algorithm). However, this is a common practice to detect outliers in psychological research (Leys, Ley, Klein, Bernard, & Likata, 2013). Importantly, our goal here is just to illustrate the presence of "extreme observations".

[5] Under normality and using the 2.5 SD criterion, 1.24% of the trials would be identified as outliers.

From these results, and in light of the aforementioned limitation of Cohen's *d* and r_{pb}/*D* effect size measures, it should be clear that there are potential problems in using either as effect size measures to reflect the difference between conditions. To illustrate, for each participant we calculated a *D* score based on the full data set (D_{full}) and a second *D* score based on the set of trials after excluding outliers (D_{excl}). That is, we implemented the *D* effect size measure within two different scoring algorithms that differed in how they deal with outliers. For illustrative purposes, we will only consider the data from one of the IRAP's four trial-types, the men-masculine trial type (and not men-feminine, women-masculine, or women-feminine trial types). Although D_{full} and D_{excl} correlate highly (r = .93), the mean absolute difference between these scores was found to equal M = .12, with SD = .11. As illustrated in Figure 1a, deviations between D_{full} and D_{excl} can be as extreme as .53. From this figure, we observe that for 5% of participants in our sample, the deviation between the two versions of *D* was larger than the standard deviation of the D_{full} scores in the sample (SD = .34). That is, the data points for 10 participants fell outside the dotted lines on the graph. Outlier data points therefore may have an unwanted influence on the scored data. This is especially problematic given that extremity is defined by arbitrary rules that may differ among researchers.

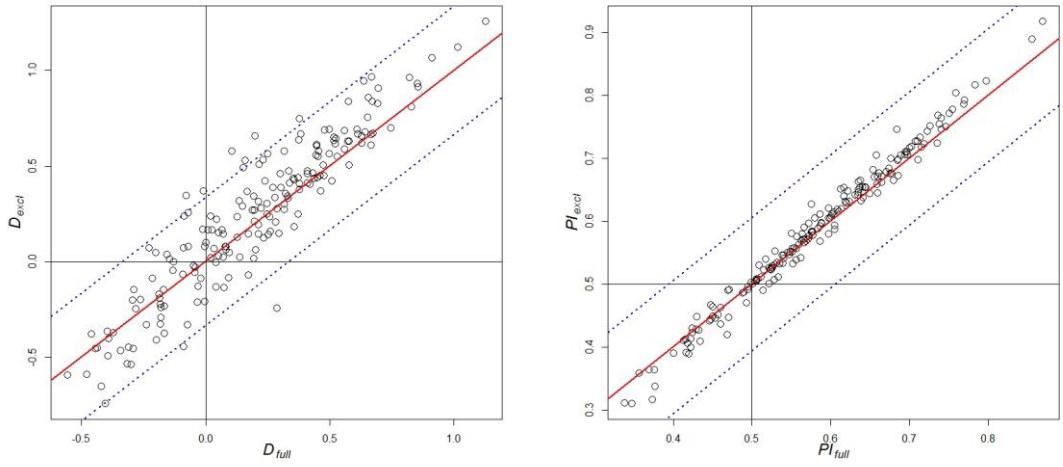


Figure 1. Left Panel: Figure 1a; Scatterplot of the obtained *D* scores using the full dataset (D_{full}) or after excluding outliers (D_{excl}). Right Panel: Figure 1b; Scatterplot of the obtained PI scores using the full dataset (PI_{full}) or after excluding outliers (PI_{excl}). Solid lines represent equal scores, dotted lines represent plus or minus one standard deviation of the sample scores (using the full datasets).

4.2 The PI: An alternative standardized effect size measure

Thas, et al., (2012) recently introduced a new class of semiparametric regression models called Probabilistic Index Models (PIMs). In the current context, a Probabilistic Index (PI) can be interpreted as the probability that a randomly selected inconsistent trial has a larger RT than a randomly selected consistent trial. As an expression of probability, the PI can therefore range between 0 and 1, where; 0 would refer to situations where RTs for all consistent trials were faster than all inconsistent trials, 1 refers to situations where RTs for all inconsistent trials are faster than all consistent trials, and 0.50 refers to situations where there is no systematic difference between the two. Importantly, the PI treats data as ordinal rather than interval, thus "faster" here refers to the fact that one reaction time (e.g., 1000) is (simply) "faster" than another (e.g., 1100), rather than being "faster by 100 ms". This is the key difference from other effect size measures that serves to minimize the influence of outliers.

In the context of the IRAP, the PI can be calculated in an easy way that immediately illustrates its interpretation (see Table 1); the reader is referred to Appendix 1 for a mathematical definition of the PI and its application to the IRAP. Suppose we observed three RTs related to consistent trials (500, 600, 700) and three RTs related to inconsistent trials (550, 650, 750). By creating the set of "pseudo-observations" (i.e. all possible pairs between consistent and inconsistent trials, in this case 3x3 = 9), we count the number of RTs faster for consistent trials compared to inconsistent trials. In this example, there are 6 pairs for which the RTs of consistent trials are faster (thus smaller). Dividing this sum by the total number of comparisons, it follows that PI = 6/9 = 0.67. We would therefore conclude that the probability that inconsistent trials have larger RTs than consistent trials was 0.67. As such, we would reformulate the original basic hypothesis of the IRAP as the probability of observing faster reaction times on trials that are consistent with a participant's behavioral history when compared to reaction times on trials that are inconsistent with that history.

To illustrate the impact of using the PI, instead of the D score, with the IRAP we calculated the PI for each participant from the gender stereotyping dataset. Here again, the PI is calculated only for the men-masculine trial type. We calculated scores from both the full data set (PI_{full}) and from the dataset after removing outliers using the same criteria as before (PI_{excl}). Note that the latter is included only to explore the influence of outliers on the PI, and

not as a recommendation that outliers should generally be defined and excluded when calculating the PI. As illustrated in Figure 1b, the mean absolute difference between these score equals M = 0.014 with SD = 0.012. More specifically, this graph illustrates that, for the PI, there were no participants whose PI_{full} and PI_{excl} scores deviated from the regression line (maximum deviation = 0.062) by more than the standard deviation of the PI_{full} score (SD = 0.105). Additionally, PI_{full} and PI_{excl} were found to correlate almost perfectly (r = .99), which was significantly higher than the correlation between D_{full} and D_{excl} (r = .93, r_{dif} = 0.06, 95%CI = [.05, .09]). As such, the PI was demonstrated to be less influenced by outlier data - and thus the arbitrariness of the rules to define outliers - than D. As we have previously discussed, this may be particularly important when working with reaction time data, in which outliers are very common.

Table 1. Calculation of the PI in a simple setting using a set of pseudo-observations for three consistent trials (500, 600, 700) and three inconsistent trials (550, 650, 750).

RT Consistent	RT Inconsistent	Inconsistent > Consistent? (if Yes = 1; If No = 0)
500	550	1
500	650	1
500	750	1
600	550	1
600	650	0
600	750	0
700	550	1
700	650	1
700	750	0
		Sum = 6 Number of pairs = 9 **PI = 6/9 = 0.67**

Lastly, to assess the presence of a linear relation between scores and general responding speed (GRS), which is, itself, often correlated with spurious variables such as age, we calculated the correlation between the absolute mean difference between consistent and inconsistent blocks and GRS (r = .55), D_{full} and GRS (r = .01) and PI_{full} and GRS (r = .01). Clearly, both D_{full} and PI_{full} reduce the unwanted correlation between the non-standardized effect size measures and the general response speed, as is typically desirable. In sum, the preceding analyses therefore suggest that the PI effect size measure is more robust to outlier data than is the D score.

4.3 An additional scoring algorithm for the IRAP: PI_{IRAP}

In the previous section we introduced an alternative standardized effect size measure for expressing the difference in performance on consistent and inconsistent trials. In this section, we will propose a new scoring algorithm making use of the PI. Readers familiar with the classic D_{IRAP} scoring algorithm will notice that the PI_{IRAP} scoring algorithm does not differ in many aspects from previously proposed measures (see for instance Barnes-Holmes et al., 2010; Hussey et al., 2015). For example, we chose to calculate one PI_{IRAP} score for each pair of test blocks and then combine them (i.e., rather than calculate a single PI_{IRAP} using all consistent blocks vs all inconsistent blocks). This is purposeful, given that the primary aim of the current article is to consider an additional effect size measure rather than its implementation within particular scoring algorithms, which is beyond the scope of the current article. We propose the following steps in calculating the PI_{IRAP} scores: (1) Use only RTs from test blocks; (2) Remove those participants with at least 10% response latencies faster than 300ms; (3) Calculate for each participant four PI_{IRAP} scores, one PI_{IRAP} for each of the four trial types (e.g., men-masculine, men-feminine, women-masculine, and women-feminine). Each PI_{IRAP} is calculated by defining a set of pseudo-observations, conditional on the block pairs. That is, all consistent trials are compared with all inconsistent trials from the same pair of test blocks. The final set combines these three different sets of pseudo-observations of the three pairs of consistent and inconsistent test blocks into one single set, and $P(Y < Y' \mid X = 0, X' = 1)$, with $X = 0$ for consistent trials and $X = 1$ for inconsistent trials, is calculated. Note that we do not exclude any outlier observations prior to the calculations.

To illustrate a number of points regarding the two algorithms, the D_{IRAP} scores and PI_{IRAP} scores obtained for the men-masculine trial type from all participants in the dataset are presented in Figure 2. To take some specific examples, a $PI_{IRAP} = 0.75$ is obtained for the men-masculine trial type for participant 1. Thus, when selecting a random consistent trial and a random inconsistent trial, it is more likely that the RTs from the inconsistent trial is larger (probability = 0.75). Participant 1's D_{IRAP} score = 0.82. Thus, the ratio between the difference in average RTs between the blocks and the variance of the pooled reaction times was 0.82. Participant 3 represents a second example, whose PI_{IRAP} score for the men-masculine trial type

= 0.32. For this participant, it is more likely that larger RTs are observed for consistent trials compared to inconsistent trials (probability = 0.68, i.e., 1.0 – 0.32). This participants D_{IRAP} score = -0.42, which is the ratio between the difference in average RTs between the blocks and the variance of the pooled reaction times.

Interestingly, the direction of the scores sometimes differs. For instance, for participant 19 it is more likely that the RTs from the inconsistent trial is larger (probability = 0.65), while the difference in average RTs between blocks is negative (D_{IRAP} score = -0.21, indicating a larger average RT for consistent trials compared to the average RT for inconsistent trials). At this point, it should also be apparent that the interpretation of the PI_{IRAP} score is therefore clearer than that of the D_{IRAP} score in terms of our original question regarding probabilistic responding speeds.

Although the direction and magnitude of individual scores might differ depending on which effect size measure is used, we do not expect large differences in the patterns observed at the group level. For instance, although substantial differences between D_{IRAP} scores and PI_{IRAP} scores are observed, for the current data set the two scores correlate highly (r = .88).

Nevertheless, due to its relative insensitivity to outlier data, the PI_{IRAP} should also demonstrate higher internal reliability. We therefore split our dataset in two halves (subsets) based on an odd/even (trial index) split and calculated Cronbach's alpha[6]. A modest improvement is observed for the PI_{IRAP} scoring algorithm (α_{PI} = .37) compared to the D_{IRAP} scoring algorithm (α_D = .29).

Finally, it is also useful to note that quantifying data using a PIM allows for the calculation of more than just a single point estimate (i.e., PI_{IRAP}). For example, we can easily obtain a 95% confidence interval for each individual score. For instance, a PI_{IRAP} = 0.75 is obtained for the men-masculine trial type for participant 1. Here, the 95% confidence interval is given by [0.59, 0.86]. Additionally, the PIM allows us to test an alternative hypothesis H_a: $PI_{IRAP} \neq 50\%$

[6] Spearman-Brown corrections, rather than Cronbach's alpha, are sometimes reported in IRAP publications, but essentially they yield the same results (Bentler, 2009).

(i.e., presence of an IRAP effect) against a null hypothesis H_0: $PI_{IRAP} = 50\%$ (i.e., no IRAP effect). In this case, we can reject the null hypothesis with $p = 0.003$. That is, one can easily determine, without the need to simulate (as would be required with the D_{IRAP} score), whether individual participants produced statistically significant PI_{IRAP} effects, as well as the magnitude and confidence interval of these effects. This may be useful for, among other things, single case designs. Relatedly, PIMs also allow the researcher to add covariates to the PI formula. Among other things, this may be useful for examining the influence of specific stimuli on the IRAP effect.

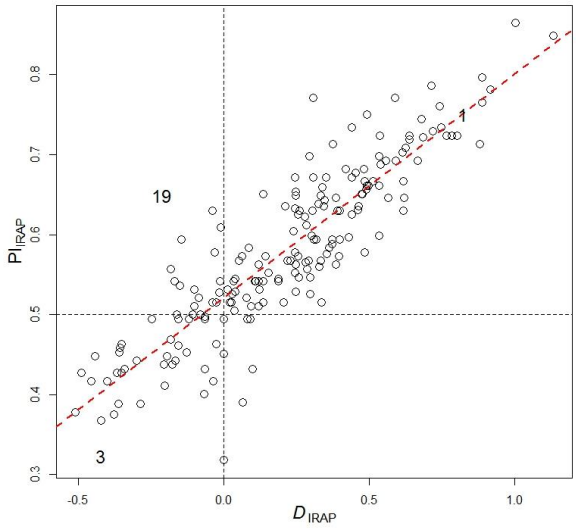


Figure 2. Scatterplot of the D_{IRAP} and PI_{IRAP} scores respectively for the entire sample. Three data points are highlighted for illustration using their participant numbers rather than a circle: "1" indicates the scores for Participant 1 and "3" indicates the scores for Participant 3, and "19" indicates the scores for Participant 19. The linear trend between D_{IRAP} and PI_{IRAP} is illustrated by a regression line.

4.4 Discussion and conclusion

In this chapter, we have introduced the PI as an alternative effect size measure to the frequently used *D* effect size measure, and then implemented it within (one possible version) of the PI_{IRAP} scoring algorithm for use with the IRAP. Although PI_{IRAP} and D_{IRAP} share similar steps, they do differ at their core: the proposed effect size measure. While the D_{IRAP} scoring algorithm defines a scaled point-biserial correlation coefficient to reflect the difference in reaction times between consistent and inconsistent trials as a proportion of the variance in all

reaction times, the PI_{IRAP} expresses this difference in performance as the probability that reaction times are higher in one context (inconsistent blocks) relative to another (consistent blocks).

As illustrated, reaction time data tends to be both heavily skewed to the right and also include outliers. Statistics, such as the sample mean and sample variance are sensitive to outliers, or as formulated by Greenwald and colleagues (1998) they "distort means and inflate variances" (p.1467). Using the sample mean and the sample variance might not be the best option, even when response latencies larger than 10000ms are omitted. In contrast, the proposed PIM-framework is much more robust to deviations from normality and to outliers. By analyzing the men-masculine trial type of the gender IRAP dataset, we have shown that (1) Substantial differences were observed between individual D_{IRAP} scores and PI_{IRAP} scores; (2) A high correlation between D_{IRAP} and PI_{IRAP} scores of the entire sample was obtained; and (3) A (moderately) higher reliability estimate was recorded.

In order to aid researchers in implementing the PI generally and PI_{IRAP} more specifically, we have included R code for a minimal implementation of the PI in Appendix II. Additionally, we produced an R Shiny web app that researchers can use to calculate PI_{IRAP} scores, which can be accessed at http://datapp.ugent.be/shiny/irap/. The source code for this app, and all code employed within the current article, can also be found on the Open Science Framework (http://osf.io/4cmsm).

We recognize that other measures for calculating effects for reaction time measures that are robust to heterogeneity and the presence of outliers have been proposed (see Richetin, Costantini, Perugini, & Schönbrodt, 2015). For instance, the Gaussian rank latency difference, or *G* score, offered by Sriram, Nosek, and Greenwald (2006) is closely related to the PI. However, the two differ with respect to their interpretation. While the PI is a direct probability, *G* scores reflect scores on a Gaussian distribution obtained by transforming fractional ranks. As such, the PI is arguably easier to interpret. In addition, given that the PI is a model-based measure, it allows researchers to calculate PI_{IRAP} scores but also confidence intervals, *p* values, and the inclusion of additional covariates (e.g., the impact of specific stimuli). In our opinion, these properties make the PI_{IRAP}, and PIM models more broadly, an interesting and highly useful choice among effect size measures that may be of use in future research.

We should reiterate that the current article has focused on the choice of effect size measure (i.e., *D* vs. PI), but has not addressed broader questions concerning other aspects of the scoring algorithm beyond the effect size measure. In contrast, Greenwald et al. (2003) made comparisons between six scoring algorithms that employ the *D* effect size measure (i.e., D_1 to D_6), but which adjust other aspects of the algorithm. Future research should therefore compare variations in PI_{IRAP} scoring algorithms that implement the PI effect size measure.

Finally, in closing it is important to recognize that it would, of course, be premature to conclude on the basis of one article, which employed only one data set and did not address the issue of predictive validity, that the PI_{IRAP} should now be used instead of the D_{IRAP}. Working out the strengths and weaknesses of specific scoring algorithms for RT measures is a complex and difficult task (e.g., Richetin, et al., 2015), and it would be unwise for researchers to adopt the PI_{IRAP} scoring method based simply on a "knee-jerk" reaction to the limited set of analyses we have presented here. Furthermore, we would advise against adopting a one-size-fits-all approach to selecting a scoring algorithm. Indeed, this may be particularly important for the IRAP because it has been used in fundamentally different ways (e.g., as a measure of implicit attitudes, as a measure of the relative strength of arbitrarily applicable relational responding, and even as a method for training and testing flexibility in relational responding). The purpose of the current work was simply to alert researchers to some of the benefits of the PI_{IRAP} relative to the D_{IRAP} effect size measure in dealing with the influence of outliers and skew effects.

References

Barnes-Holmes, D., Barnes-Holmes, Y., Hussey, I., & Luciano, C. (2016). Relational Frame Theory: Finding its historical and intellectual roots and reflecting upon its future development. In R. D. Zettle, S. C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds.), The Wiley handbook of contextual behavioral science (pp. 129–178). New York, NY: Wiley-Blackwell.

Barnes-Holmes, D., Barnes-Holmes, Y., Luciano, C., & McEnteggart, C. (2017). From the IRAP and REC model to a multi-dimensional multi-level framework for analyzing the dynamics of arbitrarily applicable responding. *Journal of Contextual Behavioral Science,*

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record*, *60*(3), 527.

Barnes-Holmes, D., Finn, M., McEnteggart, C., & Barnes-Holmes, Y. (2017). Derived stimulus relations and their role in a behavior-analytic account of human language and cognition. *The Behavior Analyst.*

Barnes-Holmes, D., Hayden, E., Barnes-Holmes, Y., & Stewart, I. (2008). The Implicit Relational Assessment Procedure (IRAP) as a response-time and event-related-potentials methodology for testing natural verbal relations: A preliminary study. *The Psychological Record, 58*(4), 497.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*(1), 137-143.

Cartwright, A., Hussey, I., Roche, B., Dunne, J., & Murphy, C. (2017). An investigation into the relationship between the gender binary and occupational discrimination using the Implicit Relational Assessment Procedure. *The Psychological Record, 67*(1), 121-130.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*, 494-509.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Cullinan, V. A., Barnes, D., & Smeets, P. M. (1998). A precursor to the relational evaluation procedure: Analyzing stimulus equivalence. *The Psychological* Record, 48(1), 121–145.

De Neve, J. (2013). *Probabilistic index models*. Unpublished doctoral dissertation. Ghent University. Faculty of Sciences, Ghent, Belgium.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin, 125,* 777-799.

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick & M.S. Clark (Eds.), *Research methods in personality and social psychology* (pp. 74-97), Newburry Park, CA: Sage.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464-1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm. *Journal of Personality and Social Psychology*, *85*, 197-216.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). Relational frame theory: A post-Skinnerian account of human language and cognition. New York: Kluwer Academic/Plenum Press.

Hughes, S., & Barnes-Holmes, D. (2016). Relational Frame Theory: The basic account. In R. D. Zettle, S. C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds.), The Wiley handbook of contextual behavioral science (pp. 129–178). New York, NY: Wiley-Blackwell.

Hussey, I., Barnes-Holmes, D., & Barnes-Holmes, Y. (2015). From Relational Frame Theory to implicit attitudes and back again: clarifying the link between RFT and IRAP research. *Current Opinion in Psychology, 2*, 11–15. http://doi.org/10.1016/j.copsyc.2014.12.009

Hussey, I., Barnes-Holmes, D., & Booth, R. (2016). Individuals with current suicidal ideation demonstrate implicit "fearlessness of death." *Journal of Behavior Therapy and Experimental Psychiatry*, *51*, 1–9. https://doi.org/10.1016/j.jbtep.2015.11.003

Hussey, I., Thompson, M., McEnteggart, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2015). Interpreting and inverting with less cursing: A guide to interpreting IRAP data. *Journal of Contextual Behavioral Science*, *4*(3), 157-162. http://doi.org/10.1016/j.jcbs.2015.05.001

Kavanagh, D., Hussey, I., McEnteggart, C., Barnes-Holmes, Y., & Barnes-Holmes, D. (2016). Using the IRAP to explore natural language statements. *Journal of Contextual Behavioral Science, 5*(4), 247–251. http://doi.org/10.1016/j.jcbs.2016.10.001

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49(4)*, 764-766. http://doi.org/10.1016/j.jesp.2013.03.013

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *11*, 361-365.

O'Toole, C., & Barnes-Holmes, D. (2009). Three chronometric indices of relational responding as predictors of performance on a brief intelligent test: the importance of relational flexibility. *The Psychological Record*, *59*, 621-640.

Richetin, J., Costantini, G., Perugini, M., & Schönbrodt, F. (2015). Should we stop looking for a better scoring algorithm for handling Implicit Association test data? Test of the role of errors, extreme latencies treatment, scoring formula, and practice trials on reliability and validity. PLoS ONE 10(6): e0129601. doi:10.1371/journal.pone.0129601

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*, 19-30. http://doi.org/10.1037/1082-989X.13.1.19

Sriram, N., Nosek, B. A., & Greenwald, A. G. (2006). Scale invariant contrasts of response latency distributions. Unpublished manuscript. http://dx.doi.org/10.2139/ssrn.2213910

Thas, O., De Neve, J., Clement, L., & Ottoy, J. P. (2012). Probabilistic index models (with Discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*, 623-671. http://doi.org/10.1111/j.1467-9868.2011.01020.x

Vahey, N. A., Nicholson, E., & Barnes-Holmes, D. (2015). A meta-analysis of criterion effects for the Implicit Relational Assessment Procedure (IRAP) in the clinical domain. *Journal of Behavior Therapy and Experimental Psychiatry*, *48*, 59-65. http://doi.org/10.1016/j.jbtep.2015.01.004

Chapter 5

On the Reliability of Implicit Measures: Current Practices and Novel Perspectives

Maarten De Schryver, Sean Hughes, Jan De Houwer, and Yves Rosseel

Abstract

During the last two decades a new class of indirect measurement procedures has emerged that has been used widely in psychological science. These procedures were developed in order to circumvent the limitations of self-reports and crack open the hidden world of 'implicit' cognition. Yet despite their popularity there seems to be no general framework that constraint (or guides) the way in which we think about the reliability of implicit measures. Instead there are far too many (subjective) researcher degrees of freedom when it comes to deciding how to assess, interpret, and even report reliability. In this paper we introduce such a framework and argue that it can be used by novel and seasoned researchers alike to estimate and interpret the reliability of their implicit measures. Our approach draws on Latent Variable Modeling and the idea of parceling in order to approximate reliability (in the sense of consistency, equivalence, stability) and test the assumptions that underlie those approximations. We close by discussing the implications of our framework for the conceptualization of reliability in particular and for implicit cognition research more generally.

5.1 Introduction

At the turn of the century, a measurement revolution swept through psychological science altering much of the methodological, theoretical, and empirical landscape. Until then researchers interested in assessing people's attitudes, beliefs, and personality characteristics most often relied on a set of *direct measurement procedures* such as semantic differential scales, feeling thermometers, and questionnaires. Although useful, these procedures were rooted in the assumption that people have both introspective access to, as well as the opportunity and motivation to accurately report on their psychological attributes. The problem, however, is that this assumption is unlikely to hold many situations, for instance, when measuring socially sensitive issues or in situations when the demand to respond in certain ways is high.

With the introduction and refinement of *indirect measurement procedures* researchers sought to circumvent these problems and crack open the hidden world of 'implicit' cognition. Broadly speaking, indirect procedures seek to (a) circumvent a person's ability to strategically control their behavior as well as (b) capture psychological processes, attributes, or content in ways that do not depend on introspective access. A vast and ever-growing library of indirect procedures have now been developed and applied to issues both in- and outside of psychology (for a recent review, see Gawronski & Payne, 2011). The most influential of these procedures include the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), evaluative and semantic priming (e.g., Fazio, 2001), the Affective Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005) and the Implicit Relational Assessment Procedure (Barnes-Holmes et al., 2006) (for more on indirect procedures see Gawronski & De Houwer, 2014; Nosek, Hawkins, & Frazier, 2011).The outcomes of indirect measurement procedures are typically referred to as implicit measures. [1]

The introduction of any new type measure in psychology sparks questions about its psychometric properties. Implicit measures were no different, with researchers raising questions about their validity (e.g., Greenwald, Poehlman, Uhlmann, & Banaji, 2009) and

[1] In-line with De Houwer and Moors (2010), we define a measurement procedure as "direct" or "indirect" based on the way in which the measurement context is arranged to capture the behavior of interest. We define the outcome derived from measurement procedures as implicit or explicit based on the automaticity features of the process by which a psychological attribute influences measurement outcomes.

reliability (e.g., Nosek, Greenwald, & Banaji, 2007). Indeed, it would not be too much of a stretch to suggest that reliability has been a persistent and thorny issue in the context of implicit cognition research. This is because the vast majority of implicit measures seem to suffer from unacceptably low levels of reliability especially when compared to their explicit counterparts (see Cunningham, Preacher, & Banaji, 2001; Fazio & Olson, 2003; Gawronski, LeBel, & Peters, 2007; Uhlmann, Pizarro, & Bloom, 2008). Although there are a number of exceptions (e.g., the IAT and AMP) reliability estimates often range from "abysmally low (Bosson, Swann, & Pennebaker, 2000) to moderate (Kawakami & Dovidio, 2001)" (p.572) and are argued to have serious knock-on effects for cumulative scientific progress (LeBel & Paunonen, 2011; but see De Schryver, Hughes, Rosseel, & De Houwer, 2015). Surprisingly, however, there is no general framework that constrains (or at least guides) the way in which we think about the reliability of implicit measures and specifies how the reliability of those measures can be determined. Instead there seems to be far too many (subjective) researcher degrees of freedom when it comes to deciding how to assess reliability, whether to even report it, and how to interpret it.

5.1.1 Looking at Reliability with a Fresh Pair of Eyes

The current paper has two goals. First, it aims to provide a quick primer for those interested in the concept of reliability and its relationship to implicit measures. We fully recognize that some readers will already be familiar with some of the statistical and psychometric concepts that will be outlined in the initial sections of in this paper. Nevertheless, we included the primer so that our ideas would be accessible also to those who are less familiar with these concepts. Based on the fact that measurement theory and practice is often ignored in doctoral programs and only 27% of the students have been found to be able to apply the methods of reliability correctly (see Aiken et al., 1990; Graham, 2006), we anticipate that many potential readers would benefit from a brief primer on the concept of reliability.

Our second (and main) goal is to introduce a general framework that can be used by novel and seasoned researchers alike to estimate and interpret the reliability of their implicit measures. If we are to have cumulative scientific progress then we need to be able to place trust in the accuracy of our measures. We also need procedures that can reproduce those

measures time and time again. As we will argue later on in this paper, both the accuracy and reproducibility of measures is closely tied to the issue of reliability. Too many researcher degrees of freedom during the data collection and reporting stages can undermine our confidence in the accuracy and replicability of findings (e.g., Simmons, Nelson, & Simonsohn, 2011). Just as researchers may (either knowingly or unknowingly) adopt practices that artificially inflate the statistical significance of their findings ('p-hacking') they might also adopt practices that inflate the reliability estimates they obtain ('r-hacking'). Given the prominent role that reliability plays in every stage of scientific evaluation it seems that we need clear guidelines and recommendations that explicate and constrain researcher degrees of freedom. With this in mind, we introduce a perspective on reliability that forces researchers to test the basic assumptions that underpin their reliability estimates and to clarify the individual and environmental factors that likely influence those estimates. Such an analysis leads to a heuristic framework for reviewing and evaluating reliability in the context of implicit cognition research.

Our paper is therefore organized in the following manner. Before we tackle reliability we first highlight a common conceptual confusion between procedures and tests (Part I). Here we argue that tests should not be equated with procedures, and that when we speak about reliability we are referring to the outcomes obtained from a particular test (taken by a specific sample) which has been administered in a wider context. In order to understand reliability we first need to understand how these three (sample, test, and context) parameters interact with one another. In Part II we shift our attention to reliability and consider how it has traditionally been understood from a (strict) classical perspective. Consistent with others (e.g., Borsboom, 2005) we argue that when one tries to interface this traditional perspective with reality it quickly starts to break down and leads to the realization that reliability can only ever be approximated in three different ways (consistency, equivalence, and stability). Each of these approximations explores a different facet of reliability and involves manipulating the test-specific, general context, and individual parameters in a certain fashion. In Part III we propose that certain conceptual problems can be circumvented by drawing on the concepts of Latent Variable Modeling (LVM) and the idea of parceling. Although parceling has been subject to some controversy (Little, Cunningham, Shahar and Widaman, 2002) it is already widely (and implicitly) used to estimate reliability in implicit cognition research (e.g., split-half). This

section will highlight the assumptions that underpin these various approximations and provides concrete methods for testing them. We then draw upon an existing dataset to provide a concrete example of these ideas (Part IV). Finally, we close by discussing the implications and consequences of our framework for reliability in particular and implicit cognition research more generally.

5.2 Part I: Clarifying the Difference between Procedures and Tests

A major impediment to communication and scientific progress is the tendency for researchers to use different terms when referring to the same phenomenon while at other times using similar terms to refer to different phenomena. This is particularly evident in the domain of implicit cognition, where tasks such as the IAT, AMP, and evaluative priming are variously labeled as implicit measures (Nosek et al., 2011), indirect measurement procedures (De Houwer, 2005), or tests (Greenwald et al., 1998) whereas the term "implicit measures" is sometimes reserved for scores obtained from those tasks (De Houwer & Moors, 2010). This terminological inconsistency can lead to conceptual confusion and theoretical misunderstanding and requires researchers to be clear in their respective terminology. In what follows we outline precisely what we mean by procedure, test, sample, context, and measure before applying these terms to the concept of reliability.

5.2.1 Measurement Procedure

We conceptualize a measurement procedure as that core property of a task that does not vary from one measurement occasion to the next. Others may refer to this as a 'paradigm'. For instance, when it comes to the IAT, researchers can systematically vary task parameters in the pursuit of their scientific goal, from the number of trials or blocks, nature of stimuli employed, or type of responses registered. In this way, the very same measurement procedure can be instantiated in a near infinite number of ways given the potential number of stimuli, responses, and other variable task parameters. What remains the same and unifies different variants of the IAT is a certain procedural consistency, namely way in which stimulus categories are mapped onto responses. More specifically, in the IAT, four stimulus categories are mapped on two responses, with two categories assigned to the first response and two to the second response,

in a way that varies across the two main phases of the task[2]. Likewise, whereas the stimulus onset asynchrony (SOA), type of trials, and other parameters may vary in an AMP, the presentations of primes and the requirement to assign a target to one of two categories remains constant across (and thus unifies) different instantiations of the AMP. Such a consistent procedural core can be identified for any procedure whether it is an IAT, AMP, or IRAP (also see De Houwer, 2003). If this core is changed in any way then the researcher has constructed a novel measurement procedure.

5.2.2 Test

A test refers to the specific ways in which the measurement environment has been arranged in order to produce an outcome. Whereas the measurement procedure is defined in terms of a core (and unchanging) property, tests represent the constellation of task parameters that are varied by the researcher in concrete studies. Amongst other things, IAT researchers can choose the number, order, and sequence of trials and blocks and AMP researchers can select the content and nature of the target and prime stimuli. Tests can be derived from the same or different measurement procedures with the intention of capturing the same, different, or related phenomena. For example, two tests can be constructed from the same measurement procedure (IAT) in order to measure two different domains (racial versus political attitudes). Yet it is also possible to construct a number of tests from the same procedure in order to examine similar domains (e.g., race IAT involving words vs. pictures). Although the latter tests involve the same procedure being used to achieve a similar goal, different task parameters were selected (and thus technically different tests were constructed). It is important to realize that each change – how trivial it might seem – results in the construction of an entirely new test, and like direct procedures, whenever a new test is constructed its psychometric qualities need to be evaluated. This practice is widely accepted in psychometric research because even small changes can, in principle, have big effects on reliability and validity. We see no reason why this practice should not upheld for tests derived from indirect measurement procedures. Just as it is inappropriate to equate the psychometric qualities of tests from one measurement

[2] Different descriptions of a measurement procedure are certainly possible. Good descriptions are those that are precise (i.e., allow one to determine whether a specific test is a variant of a specific procedure), broad (i.e., allow one to encompass many different variants of the procedure), and reveal how different procedures relate to each other and other procedures that are used in psychology.

procedure (AMP) with another (IAT), it is also inappropriate to equate the psychometric qualities of two tests from the same measurement procedure (e.g., two IATs). This is true regardless of whether the two tests are designed to assess the same phenomenon (e.g., race IAT using words versus pictures) or different phenomena (e.g., race versus self-esteem IAT).

5.2.3 Population and Sample

Besides identifying a measurement procedure and defining the parameters of one's test it is also necessary to define the population to which that test will be administered. In psychometrics, a sample is typically defined as that subset of the population that has been selected for testing. When it comes to research on implicit cognition, samples will typically be comprised of individuals, who are not static, unchangeable entities but rather organisms that constantly adapt and evolve. Individuals represent the constituent elements of a sample and can be matched in terms of, or vary along, single or multiple dimensions. In the previous section we argued that psychometric qualities are unique to a given test and that it is inappropriate to equate tests derived from the same or different measurement procedures. The exact same reasoning also holds for populations. Once a researcher has specified the population of interest and selected a representative sample from that population, she is then able to administer her test and infer the psychometric qualities of that test. These qualities only hold for the test given that population and cannot be generalized to other populations. Imagine, for instance, that an IAT designed to assess automatic attitudes to mathematics is administered to a group of students who study either psychology or mathematics. Whereas the psychology students will likely be *heterogeneous* in terms of their IAT scores, such that some individual's detest mathematics while others love it, their counterparts in the mathematics department may be a relatively more *homogenous* insofar as they consistently evaluate mathematics in a positive light. As soon as a test is administered to a sample of a previously untested population, its psychometric qualities need to be re-evaluated. Thus tests are constructed in order to examine a particular phenomenon for a given population. It is the *interaction* between a test and sample derived from that population (within a broader context) that determines the psychometric qualities of that test.

5.2.4 General Context

Even when researchers set out to construct a specific measurement environment and expose a specific sample of individuals to that environment, it is important to appreciate that this interaction will always take place within a wider context. For instance, race IATs may be taken during times of racial tension or peace, AMPs assessing religious attitudes may be administered inside a church or at a bar whereas an IAT investigating social conformity could be delivered by a person wearing informal clothing or a white lab coat. Therefore, whereas a measurement procedure is defined in terms of a core unchanging property of a task, and a test in terms of the constellation of task parameters that the researcher can vary, the general context refers to that constellation of contextual factors that are either present or manipulated during testing, but that do not comprise the test-specific context itself. In much the same way that changes in the parameters of a test or individuals that comprise a sample will influence the psychometric qualities of a test, so too will manipulations that alter the general context. Put another way, the test and population can remain the same and yet the psychometric qualities of that test can change due to variations in the wider context.

5.2.5 Measure

When a test is constructed based on a certain measurement procedure and administered to a specific sample within a broader context, some form of behavior will be observed and registered. In the context of research on implicit cognition this behavior will usually take certain forms, from the speed and accuracy of a response (Greenwald et al., 2009) to neural or physiological activity (Stanley, Phelps, & Banaji, 2008). Likewise, a wide variety of data-analytic techniques can be applied to these responses, from difference (D) score calculations to log and Gaussian transformations or even probabilistic indices (e.g., De Schryver, Hussey, De Neve, Cartwright, & Barnes-Holmes, under review). These techniques represent the set of consecutive steps that one has to follow in order to summarize the observed behavior. For instance, IAT scoring algorithms (e.g., Greenwald, Nosek, & Banaji, 2003) specify the trials to be excluded, how the tails of the distribution and error-trials should be treated, what kind of transformation should be performed and what mathematical operations should be applied.

Critically, during the construction of a test the researcher will specify the particular responses that are registered and the algorithmic parameters that are used to derive a score on the basis of those responses. As noted by De Houwer, Teige-Mocigemba, Spruyt, and Moors (2009), the selection of the response and the algorithm is an integral part of the core procedure. In line with De Houwer et al., we reserve the term "measure" for the outcome obtained whenever some data-analytic technique is applied to observed behavior. Depending on the technique involved the measure may reflect various aspects of the observed behavior (e.g. the maximum observed reaction time, the percentage of errors, the difference in median reaction time and so on)[3]. Figure 1 visualizes the nature of implicit measures and provides an overview of the relationship between procedures, tests, samples, contexts, and measures.

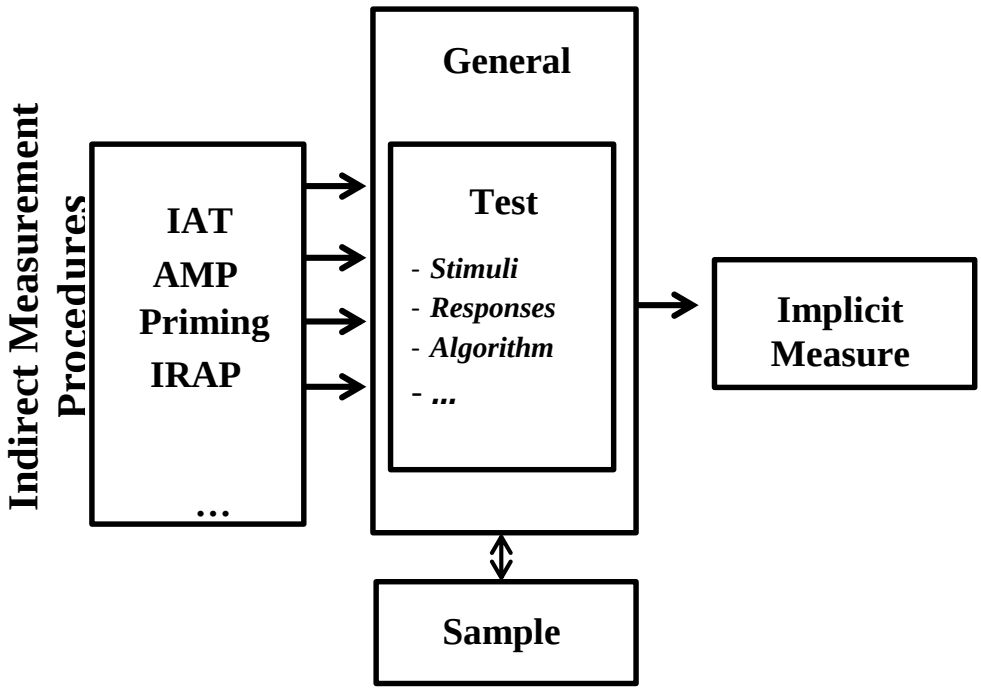


Figure 1. A visual representation of a typical measurement environment in implicit cognition research. A test is derived from a certain indirect measurement procedure and administered to a specific sample within a general context and some form of behavior is observed and registered. Based on the data-analytic techniques employed, a measure is obtained.

[3] It's worth noting that the application of a scoring algorithm to observed behavior does not imply or require that any (psychological) meaning be given to the measure. That is, a measure, defined as a summary of observed behavior, is merely the result of some set of analytic operations applied to behavior. Tests, however, are constructed with a specific theoretical or scientific goal in mind - namely - to infer knowledge about properties or attributes of individuals or groups of individuals. The relation between a to-be-measured attribute and the behavior observed on a test therefore requires some set of theoretical inferences that extend beyond the test and measure (De Houwer et al., 2009b).

5.3 Part II: Defining Reliability

5.3.1 What is Reliability?

Regardless of the procedure employed, or the specific test constructed, researchers are typically interested in the relationship between the observed score (measure) and the theoretical variable of interest. Unfortunately, any such relationship is immediately complicated by the fact that no test is capable of perfectly capturing the theoretical variable of interest and that *error* is always going to be an unavoidable companion to any observations that we make. This error typically comes in two forms: either random or systematic (see Meredith & Teresi, 2006). Random error is often viewed as trial-by-trial variability that results from unknown but non-systematic sources, such as inadvertently turning one's head away from the screen during an IAT trial or accidentally pressing a certain button during an AMP trial. Systematic error reflects trial-by-trial variability that stems from known or systematic sources such as situational or individual differences (e.g., the failure to counterbalance participant gender during a priming study on gender or systematic differences in stimuli familiarity during an IAT). Researchers are concerned with these different sources of error given that error is present in many, if not most situations, and will potentially undermine the relationship between an observed outcome and a theoretical variable. Therefore the degree of *measurement precision* that can be achieved whenever a test is administered to a given population will be determined by the extent to which error contributes to the scores obtained from that test. It is this measurement precision that researchers are referring to when they talk about the *reliability* of a test (see Borsboom, 2005; Mellenbergh, 1996).

5.3.2 Classical Test Theory and Reliability

Historically, the concept of reliability has nearly always been interpreted through the lens of Classic Test Theory (CTT) (Lord & Novick, 1968). At its core, this account makes several assumptions. The first is that each of the observed scores (Y) obtained from a test for a given population (P) can be decomposed into two independent elements: the underlying true score (T) and an error score (E) (i.e., $Y = T + E$ for P). Based on the assumptions that T and E are

independent, it can be shown that observed score variance (σ_Y^2) in the population is equal to the sum of true-score variance (σ_T^2) and error score variance (σ_E^2) (i.e., that $\sigma_Y^2 = \sigma_T^2 + \sigma_E^2$). This decompositional perspective of observed, true, and error scores reveals a key point: as the contribution of error-score to observed score variance increases the relative contribution of true-score variance to observed score variance will (by definition) decrease.

Put simply, reductions in error-score variance will lead to increases in measurement precision (for a fixed amount of true-score variance) whereas the opposite is true whenever the contribution of error-score variance grows (i.e., our measurement precision will decrease for a fixed amount of true-score variance). Therefore, from a Classical Test Theory perspective, when we speak of reliability we are referring to the ratio of true-score variance (σ_T^2) to observed score variance (σ_Y^2) and this ratio is usually written in the following way: $\rho_{YY'} = \sigma_T^2/\sigma_Y^2$, with Y' referring to a observed scores obtained from a parallel test (see below).

5.3.3 Theoretical implications versus practical applications.

Although CTT has proven highly influential, its limitations are widely known. These stem from the fact that the theory does not make any assumptions about an underlying theoretical construct that is responsible for the behavior obtained from a test (Borsboom & Mellenbergh, 2002; Sijtsma, 2009). As we just mentioned, CTT defines reliability as the ratio of true-score to observed score variance. Calculating this ratio is complicated by the fact that the true score is an unobservable quantity, as is the notion of true-score variance. We therefore need a way to make the unobservable observable. Lord and Novick (1968) demonstrated that by first constructing two parallel tests (obtaining Y and Y'), and then examining the covariance between the measures derived from those tests, true-score variance could be obtained if one assumes that participants are (a) identical with respect to each of the test administrations and that (b) the two tests are administered under the same contextual or environmental conditions. Yet, by definition, even minor modifications to the test, general context, and/or sample will yield new (true) test scores. Therefore the only parallel test that can be constructed from this perspective is the test itself, administered under identical environmental conditions, which seems only possible when the participant is somehow "brain-washed" and exposed to the test once again (see Borsboom, 2005, for an in-depth discussion).

Lord and Novick's solution may be theoretically possible, but in reality, it is practically untenable. Reliability will always remain a theoretical construct and can never be observed. This is because a precise estimation of reliability requires the general context, test-specific context, and sample to somehow remain constant across repeated test administrations. Yet as we have seen in Part I there are many different ways in which the context and sample vary from one moment to the next and the probability that all these parameters will remain fixed across time and space is practically zero. Therefore, in reality, researchers can only ever *approximate* the reliability of a test and tend to do so using one of three "proxies" known as internal consistency reliability (consistency), parallel test reliability (equivalence), and test-retest reliability (stability).

When we combine the arguments from Parts I and II an important point emerges: whenever researchers set out to approximate reliability they are (implicitly) deciding which set of (test, context, and sample) parameters they will manipulate in order to examine measurement precision. In Table 1 we provide a taxonomy that highlights which aspects of the test, sample, and context will likely vary or remain constant when these three proxies (consistency, equivalence, and stability) are approximated.

Table 1. An overview of the various approximations of reliability (consistency, equivalence and stability) and the likely elements of the general context, test-specific context and sample that will vary (v) or remain constant (c).

Reliability				
	Theoretical	Approximations		
Parameter		Internal Consistency (*Consistency*)	Parallel Tests (*Equivalence*)	Test-Retest (*Stability*)
General Context	C	C	C/V	C/V
Test-Specific Context	C	V	V	C
Stimuli	C	V	C/V	C
Responses	C	C	C/V	C
Algorithm	C	C	C/V	C
Sample	C	C	C/V	C/V
Individual	C	V	V	V

Table 1 clarifies that the three main methods of approximating reliability (consistency, stability, and equivalence) each differ in terms of the environmental parameters that vary or

remain constant. Take consistency: researchers interested in implicit cognition typically rely on consistency as an approximation of reliability whenever they assess data from a single test administered. Given that most indirect procedures are completed in a relatively short period, the general context will (likely) remain constant from one trial to another. In contrast, certain elements of the test-specific context will vary (e.g., one trial on a race IAT might present an image of a black individual while another might present a White individual) while others will remain fixed (e.g., the same class of responses are emitted during the AMP whereas the same algorithms are applied to the data from different AMP trials). Although the sample as a whole will tend to remain constant (providing attrition is low) the individuals that comprise the sample may vary in how they respond to the elements that comprise the test.

Now compare this to stability and equivalence. Implicit cognition researchers typically rely on stability whenever they want to assess the consistency of their data from one test administration to the next. They rely on equivalence whenever they want to assess the consistency of their data across tests. Although the general context may be similar across repeated test administrations the probability that it can vary is higher than when consistency is estimated within a single test administration. The test specific context will vary in a similar way to that seen when internal consistency is assessed while the individuals that comprise the sample will likely change across repeated test administrations (e.g., people may alter how they behave at Time 2 based on their experiences at Time 1).

5.4 Part III: A General Framework for Approximating Reliability

5.4.1 Latent Variable Models

Unsurprisingly, the aforementioned conceptual problems of CTT have led to alternative psychometric perspectives on reliability. One of these perspectives is based on a class of Latent Variable Models (LVM) that seek to explain different patterns of behavior in terms of some underlying theoretical construct. Before we showcase how these models address the issues of reliability we will first provide a brief introduction to LVM and related concepts.

The main aim of LVM is to explore the relationship between some unobservable or latent construct (η) and one or more observable indicators ($Y_1, Y_2, \ldots, Y_k$). Confirmatory Factor Analysis (CFA) is a well-known example of LVM and is often used to explore how the latent construct is linearly related to its various indicators. In CFA the relationship between a latent construct and its indicators is often modeled using a regression equation and is defined by three (to-be-estimated) sets of parameters: the intercept (a_i), slope or factor 'loading' (b_i) and error scores (e_i) (see Figure 2). Thus CFA can be mathematically defined as $Y_i = a_i + b_i\eta + e_i$, for indicator ($Y_i$) (Jöreskog, 1971; Raykov, 2004; Brown, 2006; Graham, 2006).

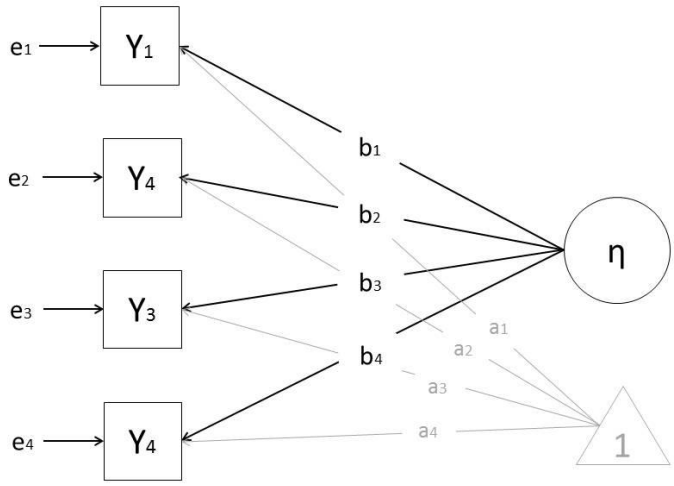


Figure 2. A visual illustration of a CFA model where the relationship between an unobservable or latent variable (η) and its observable indicators (Y_i) is determined by three to-be-estimated parameters: the intercept (a_i), slope (b_i), and error term (e_i).

Three different types of CFA models are typically constructed and each differs in the specific assumptions it makes about the above parameters (a_i, b_i and e_i). First, *congeneric* models allow all of the parameters to vary under the assumption that the indicators measure the same latent construct (Brown, 2006). Second, (*essentially*) *tau-equivalent* models take a different approach, constraining the factor loadings (b_i) to be equal, while allowing the other model parameters to vary[4]. Third, *parallel* models take the most restrictive route by forcing

[4] An essentially tau equivalent model differs from a tau equivalent model by allowing the intercept a_i to vary across components. While both models assume that the component measures relate in a similar way to the latent construct η, essentially tau equivalent models allow for different means across the

the factor loadings and the error variances to be equal. In this way a parallel model is nested in a tau-equivalent model and a tau-equivalent model is nested in a congeneric model (i.e., all three models have the same structure but differ in the parameters that they fix or let vary). To test if the model actually meets its criteria (i.e., that certain parameters are equal to one another) model comparison tests are conducted (e.g. a χ^2 - difference test; see Brown, 2006).

5.4.2 Indicators in Implicit Cognition Research.

The relationship between the latent construct and its indicators is pivotal for CFA. Hence, it is important to clearly specify what an indicator is. By definition, an *indicator* has to fulfill two basic conditions: it must vary systematically with changes in the latent construct and increase or decrease monotonically with the latent construct (Lord & Novick, 1968). In other words, when higher (or lower) scores are observed on the indicator this must be caused by increases (or decreases) in the latent construct's values. This relatively simple story is immediately complicated when it comes to indirect procedures like the IAT, IRAP, and certain types of priming. Due to the ways in which these tasks are constructed, the responses observed in one condition (e.g. those obtained from congruent or incongruent blocks of trials) cannot be considered as indicators of the construct of interest. Although these responses are certainly related to the latent construct they often violate the aforementioned conditions (e.g., they do not always vary monotonically with the construct of interest). Several authors have acknowledged this finding in the literature. For instance, Nosek and Sriram (2007) argued that the responses obtained from congruent or incongruent blocks on an IAT cannot be treated as direct indicators of the underlying construct and that, just like the Stroop task, it is only by contrasting these two conditions that one can obtain an indicator in indirect procedures such as the IAT. Therefore when we speak of indicators we are referring to the application of some mathematical operation to the behavior obtained from a test. This is precisely the same way in which we defined a measure in Part I.

The process of selecting an indicator, that is (certain) elements of a test to which mathematical operations will be applied in order to generate a measure, can be defined as

measures. The component measure and the (essentially) tau-equivalent model make no restriction with respect to the error variances. In both models, different measurement precision can be obtained for the different components.

parceling. It is through this process of parceling that one obtains parcels. We define a parcel as *an indicator for which the same mathematical operations that would be applied to the entire test have been applied to a subset of elements of the test.*[5] There appear to be two main ways of parceling one's data. On the one hand, a distributed parceling strategy can be used wherein items are either randomly or sequentially assigned to a parcel. The simplest version of this involves dividing one's data into two halves based on an odd/even split (LeBel & Paunonen, 2011), although researchers have also carved their data into three (Cunningham et al., 2001), four (Mierke, & Klauer, 2003) or even more parcels using a similar strategy (Schmitz, Teige-Mocigemba, Voss, & Klauer, 2013). On the other hand, a homogenous parceling strategy can be applied wherein the items that comprise the parcel are in some way related to one another, and thus form a homogenous group (Marsh, Lüdtke, Nagengast, Morin, & Von Davier, 2013).

Two points are worth noting here. On the one hand, there are many researcher degrees of freedom when it comes to parceling: one not only has to decide how many parcels should be created but also what parceling strategy should be used (homogenous vs. distributed). Specifying a priori how and why one will parcel their data places constraints on this process and thus limits those degrees of freedom. On the other hand, certain parceling strategies have the potential to camouflage unwanted sources of variance. A popular idea in the implicit cognition literature is that the trials that make up a certain phase of a test (e.g., those in the congruent block) are all equivalent and that the responses to any one trial within this phase will be similar to the next (i.e., there is a high exchangeability of trials within a block). Therefore the variation in responding to these trials (at least within participants) is thought to represent random error variance (Blanton, Jaccard, & Burrows, 2015). If this idea is correct, then parceling one's data in different ways (e.g., either randomly, sequentially, or even homogenously) should yield similar outcomes.

Imagine, for instance, that a researcher decides to assign trials from a race IAT to different parcels based on their content. She assigns Black exemplars to a 'Black' parcel, White exemplars to a 'White parcel', positive items to a 'positive parcel' and negative items to a

[5] Note that in the LVM literature parcels are typically defined as "indicators comprised of the sum (or average) of two or more items, responses, or behaviors" (Little, Cunningham, Shahar, & Wideman, 2002, p.152). Yet many mathematical operations can be applied to elements of a test (above and beyond simple summation or averaging) when constructing parcels.

'negative' parcel. If the above exchangeability assumption holds then these content-based parcels should all be equivalent - or at least tau equivalent. Yet there is no guarantee that this assumption will hold. For instance, there could be some systematic confound that applies to certain test-elements but not others. Take the above example: participants may be more familiar (thus able to respond quicker) to in-group exemplars and be less familiar with (and thus slower to respond to) out-group exemplars. They may also be better able to process certain types of stimuli (e.g., images of faces) compared to others (e.g., names). The test-specific context might also influence responding in other ways (e.g., images presented on black vs. white backgrounds might facilitate or undermine how readily people identify Black and White faces). Now when a distributed parceling strategy is applied the above systematic variance is potentially 'hidden' or 'spread' across the various parcels - and as such - the latent variable might capture that systematic variance and thus overestimate consistency (see Little, Rhemtulla, Gibson, & Schoemann, 2013; Maul, 2013). Yet a content-based strategy avoids this problem by forcing the systematic variance to remain unique to each parcel. Therefore we recommend that researchers stay vigilant to such unwanted sources of variance that may contaminate their consistency estimates. One way to do this is to use a homogenous (content-based) parceling strategy that at least allows for these systematic sources of variance to be identified.

5.4.3 Latent Variable Models and Reliability

So, taking a step back, what does all of this have to do with reliability? As we outlined above, CTT makes certain theoretical assumptions about reliability that seem rather difficult (if impossible) to instantiate in reality. Faced with the need to estimate the precision of their measurements, researchers have often ignored these theoretical assumptions and instead tried to approximate the degree to which their data was internally consistent, equivalent, or stable from one measurement occasion to the next. We also argued that each of these approximations is basically interested in answering a different question and will therefore proceed from the position that certain aspects of the general context, test, and sample should be fixed or allowed to vary. Therefore we need a way of conceptualizing these different approximations and of putting their respective assumptions to the test.

Latent Variable Models can do just this. They not only address many of the theoretical problems outlined above but, perhaps more importantly, they provide an overarching method for conceptualizing the different approximations and for testing their respective assumptions. Whereas CTT (strictly speaking) considers different parts of a test, or different tests to measure different psychological concepts, LVM allows us to assume that performance on different implicit measures can actually be underpinned by the same latent construct. In this way LVM go "beyond classical test theory in that they attempt to construct a hypothesis about the data-generating mechanism in which the attribute is explicitly represented as a latent variable" (Borsboom, 2005, p.49). LVM allows us to explore how parcels are related to a latent variable by comparing different models (e.g., does a parallel or tau-equivalent model fit the data adequately?) One of the main advantages of thinking in this way is that it allows researchers to examine the assumptions that underpin the various approximations of reliability. For example, by using CFA the specific assumptions that different approximations of reliability make, can be tested by examining the covariation between the different parts of a test (Yang & Green, 2011). In what follows we will unpack these points in greater detail. Specifically, we will describe the three reliability approximations (consistency, stability, and equivalence), indicate what their underlying assumptions are, as well as the consequences that follow when their assumptions are violated. We will then show how CFA can be used to explore these assumptions and to approximate reliability.

5.4.4 Consistency as an Approximation of Reliability

We define *consistency* as the interrelatedness of *a priori* constructed parcels of a test.[6] In the next section we will show how the parameter estimates obtained from a CFA model can be used to generate a consistency estimate known as coefficient 'omega' (ω, see McDonald, 1978, 1999; Raykov, 2001, 2004; Bentler, 2009). We then argue that two of the most common ways of approximating consistency in the implicit cognition literature (s*plit-half reliability* and *Cronbach's alpha*) are just two instances of this coefficient omega (Raykov, 2004). The value of the (omega) approach lies in the fact that it cannot only generate split-half and Cronbach's

[6] Note that the concept of internal consistency can also be defined in other ways. For instance, one can view it as a measure of *homogeneity* (i.e. the degree of unidimensionality of the items; for a discussion see Sijtsma, 2006; Clark & Watson, 2009; Yang & Green, 2011).

alpha but also yields a reliability estimate in situations where the assumptions underpinning the above do not hold (i.e., for congeneric models).

Coefficient Omega. Although little used in the implicit cognition literature estimating coefficient omega via CFA is a **"***widely applicable latent variable modeling approach to point and interval estimation*" of consistency (Raykov, 2004, p.324) that can be mathematically defined as $\rho_{YY'} = \frac{(\sum_{i=1}^{k} b_i)^2}{(\sum_{i=1}^{k} b_i)^2 + \sum_{i=1}^{k} \theta_{ii}}$, where k is the number of parcels, b_i is the factor loading, and $\theta_{ii} = \mathrm{Var}(e_i)$ is the error variance of the i^{th} parcel[7]. The advantage of this approach is that one can easily estimate consistency in many different ways using a single formula. For instance, imagine that you have two parcels (k = 2), and that your factor loadings ($b_1 = b_2$) and error terms ($e_1 = e_2$) are equal to one another. The solution yielded by the above formula will be the same as what one obtains from the split-half reliability formula. Now imagine that you have k parcels and that the factors loadings of those parcels ($b_1 = b_2 = \ldots b_k$) are equal. This same formula will yield an outcome similar to that obtained from the Cronbach's alpha formula given the same number of parcels. Interestingly, in situations where factor loadings (or factor loadings and error terms) are not equal, the omega formula can still be used to generate an unbiased consistency estimate given a specific set of indicators. Although split-half reliability and Cronbach's alpha represent the most popular means of determining the consistency of one's data, the assumptions that underlie these estimates are rarely tested (or at least rarely reported). For instance, when one calculates split-half reliability using the Spearman-Brown correction (which can be mathematically defined as $\hat{r}_{sb} = \frac{2\hat{r}}{1+\hat{r}}$) they are assuming that the two parcels they have created are parallel in nature (Eisinga, Grotenhuis, & Pelzer, 2013). Recall that from a CFA perspective this means that the measures obtained from those parcels are thought to reflect the same underling construct, that the factor loadings are identical, that the error variances are equal, and that the errors do not correlate. Likewise, when we use Cronbach's alpha, we assume that the parcels are (essentially) tau-equivalent. That is, we make a similar set of assumptions as above except that the condition of equal error variances is relaxed (Novick & Lewis, 1967; Lord & Novick, 1968; Yang & Green, 2011; Eisinga et al., 2013). Whenever these assumptions are violated we end up with problematic consistency

[7] Note that this version of the formula only holds for uncorrelated errors. For correlated errors we refer the reader to Raykov (2004, p.304).

estimates that may be over- or underestimated. Put another way, even in situations where the underlying measurement model is not parallel (which is required for split-half reliability[8]) or tau-equivalent (which is required for Cronbach's alpha) but simply congeneric, researchers can still use the above formula to estimate consistency.

The identification of the measurement model is thus an important first step before approximating reliability using the CFA methodology. Yang and Green (2011) argue that a congeneric model should first be fit to the data so that each parameter can be freely estimated. A tau-equivalent model should then be explored, and if it fits the data equally well, then it should be preferred over the congeneric model. If it does not then the congeneric model should be accepted. If the tau-equivalent model fits the data equally well then it should then be compared with a parallel model (and the later only accepted when it also fits the data).

5.4.5 Equivalence as an Approximation of Reliability

We define *equivalence* as the interrelatedness of measures obtained from two or more tests. When reliability is approximated in this way the researcher assumes that the scores obtained from two independent administrations of the same test reflects the same construct. If this is true then correlations between test scores can be considered as a direct estimate of reliability. This presents an immediate (practical) problem of constructing tests that are actually parallel to one another: in a strict sense, two tests can be considered *parallel*, if and only if, they are *exactly equivalent*. Both tests should not only measure the same (underlying) construct but also do so with the same degree of precision (i.e., observed score means, variances, and co-variances of different parts of the tests should be equal to one another).

While the conditions of parallelism can easily be met in the context of a thought experiment they are difficult, if not impossible, to realize in the real world. Consider studies that have set out to measure the same construct using "parallel tests". For instance, Dasgupta, McGhee, Greenwald, and Banaji (2000) assessed implicit racial preferences using IATs with

[8] Some authors have avoided choosing a specific parceling strategy by adopting a bootstrap procedure. This involves first splitting the data into two equal halves, calculating a correlation coefficient, and then repeating this process many times (e.g., Ravenzwaaij, van der Maas, & Wagenmakers, 2011). The mean correlation coefficient (Spearman-Brown corrected) they obtain is then used as a consistency estimate

pictorial or name stimuli. Likewise, Asendorpf, Banse, & Mücke, (2002) examined implicit shyness while Greenwald et al. (1998) examined racial evaluations using two IATs with different target stimuli. According to our framework, the authors of these various studies have manipulated aspects of the test-specific context. For instance, they have varied the nature of the stimuli while fixing the type of responses registered and algorithms applied to the data. Although their samples always remained constant across testing the individuals that comprise those samples likely differed from one test administration to the next. Whenever researchers manipulate aspects of the test-specific context, general-context, and/or the individual they produce tests that differ from one another in some way, and therefore, they need to determine if those tests are parallel as initially assumed.

Yet as is the case with consistency, the assumptions underpinning equivalence estimates are rarely - if ever - tested or reported in the literature (i.e., same latent construct and same degree of measurement precision). Again these assumptions can easily be assessed via CFA by identifying the underlying measurement model. For instance, one could construct four parcels (e.g., based on content) for one test and then do so again for the second test. If the conditions of parallelism hold then a model where both factor loadings and error variances are equal among related content parcels , should fit the data. In conditions where the model does not fit the data researchers cannot claim that the two tests are parallel - and by definition - the observed correlation coefficient between test scores cannot be considered as a good approximation of reliability.

5.4.6 Stability as an Approximation of Reliability

We define *stability* as the interrelatedness of measures obtained from the same test that has been administered two or more times. Stability and equivalence bear many similarities to one another: in each case a test is administered at least two times and the correlation between test scores are assumed to reflect the same latent construct. From our perspective, the key difference between equivalence and stability is the extent to which they involve manipulating aspects of the test-specific context. Whereas equivalence does involve manipulating the test-specific context (because different tests are used) stability does not (because the same test is used). Yet in both cases the general-context and the individual may vary given that both

approximations involve repeatedly administering tests across time (see Table 1). The extent of this variation may depend on the time lag between test administrations.

The assumptions underpinning stability are the same also those underpinning equivalence and can therefore be tested in the same way. For instance, one could construct two parcels for one test and another two parcels for a second test. If the conditions of parallelism hold then a model where the factor loadings are equal, and error variances are equal, should fit the data. In conditions where the model does not fit the data researchers cannot claim that test scores are stable - and by definition - the observed correlation coefficient between test scores cannot be considered a good approximation of reliability.

5.5 Part IV: An Empirical Example

We have just introduced a LVM approach to the approximation of reliability. This approach may be new for many researchers interested in the study of implicit cognition. Even those who are acquainted with a LVM approach they might never have applied it to the topic of reliability. Others may have been aware of both LVM and its links to reliability but never approximated the latter using the former because access to relevant software was either unavailable or too costly. As such, a short example of this approach in action will serve as a tutorial on the basic ideas outlined here. It will also show how a CFA model can be fit to data in order to generate consistency estimates using the free and open-source software known as R (R Development Core Team, 2016) and the add-on package known as lavaan (Rosseel, 2012).[9]

At the same time, this section will also showcase how the choices one makes when approximating consistency have real and tangible consequences for the types of estimates one obtains. This will be achieved by drawing on a data set that was originally reported by Bar-Anan and Nosek (2014). Although these authors administered several indirect procedures in different domains (i.e., the so-called "Attitudes 3.0" data set), we will only focus on the data of those who completed a race IAT. In this IAT the target labels were *White People* and *Black*

[9] For readers interested in learning more about CFA we recommend Brown (2006) whereas those looking to learn more about fitting CFA models using R can read Finch and French (2015). Although our example focuses on (internal) consistency it could easily be adapted for stability and equivalence as well.

People while the target stimuli were six pictures of white people and six pictures of black people. The attribute labels were *Good* and *Bad* while the attribute stimuli were six positive and six negative words. The authors calculated IAT scores using the D2 scoring algorithm. They then estimated the internal consistency of their data (using the Cronbach's alpha formula) by constructing three parcels based on a distributed parceling strategy wherein items were sequentially assigned to three different subsets. In this way they obtained an internal consistency estimate of .86.

In what follows, we first approximate reliability via the "traditional" approaches used in implicit cognition research (i.e., split-half reliability and Cronbach's alpha). We then apply the CFA-methodology using the two-step method recommend by Yang and Green (2011) (i.e., where the CFA model is first assessed and then coefficient omega estimated). Although Bar-Anan and Nosek's sample comprised White and Black participants, we apply the CFA approach in three ways: to a sub-sample of White participants (N = 161), a sub-sample of Black participants (N = 161), and a combination of the two (N = 322). Doing so will show how consistency estimates can fluctuate depending on the parceling strategies (and samples) involved, and thus why researchers need to be aware of these issues (see Appendix for R code that will allow the reader to reproduce the results of the combined sample).

Split-half reliability. When researchers have the aim to determine the consistency of an IAT based on a certain IAT data set, they typically rely on either split-half or Cronbach alpha. Take split-half reliability: by splitting the test into two parcels of equal length we obtain two measures. Given that the number of items within each parcel is half that of the original test the correlation coefficient will tend to underestimate reliability. In order to correct for this the coefficient is typically adjusted using the Spearman-Brown prophecy formula, leading to what is known as stepped-up reliability (Lord & Novick, 1968). In the implicit cognition literature split-half reliability is usually calculated by creating two different parcels based on the index number of the trial (i.e., whether it was odd or even numbered; see Glashouwer, Smulders, de Jong, Roefs, & Wiers, 2013; Krause, Back, Egloff, & Schmukle, 2011; Schmukle, & Egloff, 2006). The popularity of this (distributed) parceling strategy may stem from the idea that it leads to two different tests that are assumed to be parallel in nature. Applying this approach to

Bar-Anan and Nosek's race IAT yields a reliability coefficient that is higher for the White participants (0.79) and combined sample (0.80) than the Black participants (0.68).

Cronbach's alpha. One could also estimate consistency by creating three or more parcels (using either a distributed or homogenous parceling strategy) and then calculating Cronbach's alpha. In practice there are often few limitations placed on researchers at this analytic stage (i.e., there are many researcher degrees of freedom). For instance, Bar-Anan and Nosek (2014) constructed three parcels based on a distributed parceling strategy (without any clear motivation for why they did so). But they could have equally constructed four, five...or any number of parcels and used either of the above strategies. That said, there may be logical reasons for choosing one strategy over another.

It is worth noting here that tests are comprised of different elements could share some systematic variance that is 'reliable' but unrelated to the latent construct of interest. This type of variance might lead to an overestimation of consistency especially when a distributed parceling strategy is employed. To illustrate, imagine that we adopt two strategies. The first is a homogenous parceling strategy where trials were assigned to four different parcels based on their respective content: with Black stimuli assigned to a Black parcel, white stimuli to a White parcel, positive adjectives to a Positive parcel, and negative adjectives to a Negative parcel. In the second case we constructed four parcels via a distributed parceling strategy, with trials assigned sequentially based on their index number. Calculating Cronbach's alpha for the homogeneous parcels revealed that data from White participants (0.78) and the combined sample (0.80) was once again more internally consistent than from the Black participants (0.69). As expected, the consistency estimates for the distributed parcel data were generally higher, and the same pattern was observed (White participants (0.85); combined sample (0.86); Black participants (0.77)). However, there are no guarantees that the homogenous parcels will meet the assumptions underlying Cronbach's alpha and might produce biased consistency estimates. We therefore need a way to test these assumptions, and as discussed above, this can be achieved using CFA.

Coefficient omega. Applying this logic to the homogenous data from above means that a congeneric model (one for each of our three samples) should be constructed with the parcels as indicators and one latent variable. To identify the scale of the latent variable its variance

was set to 1.00. The congeneric model seemed to fit the data well in all three cases[10]. When these congeneric were then compared with tau-equivalent models, the former were found to fit the data better than the latter, which is indicated by highly significant *p*-values. Moreover, bad fit indices were obtained for the tau-equivalent models (see Table Y for factor loadings and error variances of the congeneric models). The parameter estimates of the congeneric models were then used to calculate coefficient omega – one for each of our samples. This can be achieved by using the lavaan syntax to define the coefficient. One advantage of the lavaan syntax is that it also provides standard errors which can be used to calculate a 95% confidence interval ($\hat{r} \pm 1.96$ SE). It is also possible to use a bootstrap-based confidence interval with the lavaan method. Calculating omega for the homogeneous parcels revealed that data from White participants (0.80) and the combined sample (0.82) was once again more internally consistent than from the Black participants (0.72) (see Table 2).

Table 2. Consistency estimates for the White participants, Black participants, and combined sample from Bar-Anan and Nosek (2014). Four estimates are provided for each sample: Split-half reliability (with Spearman-Brown corrections), Cronbach's alpha (distributed parceling strategy; Distr), Cronbach's alpha (homogenous parceling strategy; Hom), and coefficient omega.

	Combined Sample (N = 322)	**White Sample (N = 161)**	**Black Sample (N = 161)**
Split-half	.80	.79	.68
alpha (Distr)	.86	.85	.77
alpha (Hom)	.80	.78	.69
Omega	.82 (SE = 0.017) 95 CI={.79-.85}	.80 (SE = 0.031) 95 CI={.74-.86}	.72 (SE = 0.038) 95 CI={.65-.79}

Given that only a congeneric model fits the homogenous data the necessary condition of tau-equivalence does not hold - a condition upon which the use of Cronbach's alpha rests - and thus this consistency estimate will likely be biased. Indeed, as Table 2 shows, the use of Cronbach's alpha (for the homogeneous data) leads to an underestimate of internal consistency. However, consistency estimates are still lower compared to the estimates based on the distributed parcels. Table 3 also shows that there is a large difference in the factor loadings for picture (Black and White) compared to the word items (Positive and Negative) as

[10] As rules of thumb we use the following criteria for good model fits: p-value of χ^2-test > .05; RMSEA < .06; CFI > = .95 (for more see Brown, 2006). For the combined sample model fits were $\chi^2(2) = 0.15$, p=.93, CFI=1.00, RMSEA = 0.00. For the White participants model fits were were $\chi^2(2)$ 2.75, p=.25, CFI=1.00, RMSEA = 0.05. For the Black participants model fits were were $\chi^2(2) = 3.41$, p=.18, CFI=0.99, RMSEA = 0.07.

well as for items from the Black parcel compared to Positive parcel (both for White participants and Black participants), questioning the assumption of exchangeability between items.

Table 3. Factor loadings and error terms obtained from the congeneric model with four homogenous parcels (White parcel (WP); Black parcel (BP); Negative parcel (NP) and Positive parcel (PP) for the White participants, Black participants, and combined sample from Bar-Anan and Nosek (2014).

	WP	**BP**	**NP**	**PP**
Total sample (N=322)	0.60 (0.64)	0.50 (0.75)	0.88 (0.22)	0.84 (0.29)
White sample (N = 161)	0.55 (0.70)	0.49 (0.76)	0.89 (0.21)	0.85 (0.28)
Black sample (N = 161)	0.47 (0.78)	0.36 (0.87)	0.83 (0.31)	0.73 (0.46)

The above demonstrates how consistency estimates can vary as a function of the sample and parceling strategies (distributed vs. homogenous) involved (see Table 2). Consistency estimates were higher for White participants and the combined sample relative to Black participants, and (as expected) a distributed parceling strategy produced larger estimates than a homogenous one. In this case, some systematic (error) variance – and thus variance unrelated to the construct of interest - seems to be captured by the latent variable. The results also demonstrate the importance of identifying the measurement model: by applying the Cronbach alpha formula (without testing the condition of tau-equivalence) researchers might obtain biased consistency estimates.

5.6 Part V: Conclusion

In the course of designing experiments as well as collecting and analyzing data, researchers face many decisions: they need to figure out how they should structure the test and general-context, what sample they should select for testing, what data they should collect and how they should analyze that data. They are afforded many degrees of freedom during this decision making process and even more whenever it is hidden from view or they fail to specify it on an *a priori* basis. This can lead to problematic situations in which "researchers explore various analytic alternatives to search for a combination that yields 'statistical significance' and to then report only what 'worked'" (Simmons et al., 2011, p.1359).

This is also true for reliability. Every researcher is faced with multiple decisions when attempting to determine the precision of their measurement outcomes; from the reliability approximations they select (consistency, stability, equivalence), to how the data should be parceled, how many parcels they should create, and what should be assigned to those parcels (homogenous vs. heterogeneous). In the absence of clear guidelines and standards for approximating reliability the degrees of freedom afforded to researchers are high and their reliability-related decisions they make can be hidden or simply ignored. We argue that this state of affairs reflects current practice within the implicit cognition literature where the motivation for, and assumptions underpinning reliability, are rarely if ever explicated or tested. This leads to a situation where researchers can effectively 'cherry-pick' those reliability estimates that presents the most favorable impression of their implicit measures.

In the current paper we attempted to address the above problems by providing a framework for approximating reliability. We first argued that any discussion about reliability has to begin from a position of conceptual clarity. We highlighted the difference between a test and procedure and showed how implicit measures represent *outcomes* that are obtained when a constellation of test-specific elements are administered to a specific sample of individuals who complete that test within a general context. Thus when we speak about the reliability of a test we always have to consider the interaction between these three elements (sample, test, context). Second, we offered a framework that highlights the likely elements of the general context, test-specific context, and sample that can remain constant or vary. Third, we provided a LVM perspective on reliability providing a clear way of approximating reliability (consistency, equivalence, stability) and testing the assumptions that underpin those approximations. By comparing conventional (split-half and Cronbach's alpha) and more recent ways of approximating reliability (omega using the CFA approach), by highlighting the concept of parceling, and by applying the above to real-world data, we then showed that the way in which one manipulate these elements can have an impact on the consistency, stability, or equivalence estimates obtained, demonstrating the need for a clear set of standards in this area.

To conclude, when it comes to reliability we recommend that the following practices be adopted: authors should (a) motivate why they decided to approximate reliability in a given way before data collection and report this rule in their article, (b) test the assumptions that

underpin the reliability approximation they selected, and if these assumptions fail to hold, then (c) apply an appropriate correction or alternative approximation. We also encourage reviewers and editors to ensure that authors follow these recommendations. Doing so requires relatively little from all parties but will ensure the further improvement of transparent and good scientific practices (and thus cumulative scientific progress). It seems likely that high researcher degrees of freedom are partially responsible for the so-called replication crisis in psychological science as well as the historical issues surrounding the reliability of implicit measures. We hope our framework and suggestions help address both.

References

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., Roediger III., H. L., Scarr, S., ... & Sherman, S. J. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist*, *45*(6), 721.

Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: the case of shy behavior. *Journal of personality and social psychology*, *83*(2), 380.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior research methods*, *46*(3), 668-688.

Barnes-Holmes, D., Barnes-Holmes, Y., Power, P., Hayden, E., Milne, R., & Stewart, I. (2006). Do you really know what you believe? Developing the Implicit Relational Assessment Procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist*, *32*(7), 169-177.

Bentler, P. A. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*. doi:10.1007/s11336-008-9100-1

Blanton, H., Jaccard, J., & Burrows, C. N. (2015). Implications of the implicit association test D-transformation for psychological assessment. *Assessment*, *22*(4), 429-440.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, *30*(6), 505-514.

Bosson, J. K., Swann Jr., W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited?. *Journal of personality and social psychology*, *79*(4), 631.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Publications.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological science*, *12*(2), 163-170.

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2000). Automatic preference for White Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology*, *36*(3), 316-328.

De Houwer, J. (2005). What are implicit measures and indirect measures of attitude. *Social Psychology Review*, *7*(1), 18-20.

De Houwer, J., & Moors, A. (2010). Implicit measures: Similarities and differences. In B. Gawronski & K. B. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 176-193). New York: Guilford Press.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological bulletin, 135*(3), 347.

De Schryver, M., Hussey, I., De Neve, J., Cartwright, A., & Barnes-Holmes, D. (under review). The PI_{IRAP}: An alternative scoring algorithm for the IRAP using a more robust effect size measure. http://osf.io/4cmsm

De Schryver, M., Hughes, S., Rosseel, Y., & De Houwer, J. (2015). Unreliable Yet Still Replicable: A Comment on LeBel and Paunonen (2011). *Frontiers in psychology*, *6*.

Eisinga, R., Grotenhuis, M. T., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown?. *International journal of public health*, 1-6.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition & Emotion*, *15*, 115-141

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual review of psychology*, *54*(1), 297-327.

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. *Handbook of research methods in social and personality psychology*, *2*.

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us?: Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science*, *2*(2), 181-193.

Gawronski, B., & Payne, B. K. (Eds.). (2011). *Handbook of implicit social cognition: Measurement, theory, and applications*. Guilford Press.

Glashouwer, K. A., Smulders, F. T., de Jong, P. J., Roefs, A., & Wiers, R. W. (2013). Measuring automatic associations: Validation of algorithms for the Implicit Association Test (IAT) in a laboratory setting. *Journal of Behavior Therapy and Experimental Psychiatry*, *44*(1), 105-113.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability what they are and how to use them. *Educational and Psychological Measurement*, *66*(6), 930-944.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464-1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of personality and social psychology*, *85*(2), 197.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of personality and social psychology*, *97*(1), 17.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109-133.

Kawakami, K., & Dovidio, J. F. (2001). The reliability of implicit stereotyping. *Personality and Social Psychology Bulletin*, *27*(2), 212-225.

Krause, S., Back, M. D., Egloff, B., & Schmukle, S. C. (2011). Reliability of implicit self-esteem measures revisited. *European Journal of Personality*, *25*(3), 239-251.

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin*, *37*(4), 570-583.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural equation modeling*, *9*(2), 151-173.

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*(3), 285.

Lord, F. M., Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Mierke, J., & Klauer, K. C. (2003). Method-specific variance in the implicit association test. *Journal of personality and social psychology*, *85*(6), 1180.

Marsh, H., Lüdtke, O., Nagengast, B., Morin, A. J. S., & Von Davier, M. (2013) Why item parcels are (almost) never appropriate: two wrongs do not make a right-camouflaging misspecification with item parcels in CFA models. Psychological Methods, *18*, 257-284. DOI: 10.1037/a0032773

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology,V4, Article 169*, 1-13.

McDonald, R. P. (1978). Generalizability in factorable domains: "domain validity and generalizability": 1. Educational and Psychological Measurement, 38(1), 75–79.

McDonald, R. P. (1999). Test theory: A unified treatment. Hillsdale: Erlbaum

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*(3), 293.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical care*, *44(11)*, S69-S77.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. *Automatic processes in social thinking and behavior*, 265-292.

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in cognitive sciences*, *15*(4), 152-159.

Nosek, B. A., & Sriram, N. (2007). Faulty assumptions: a comment on Blanton, Jaccard, Gonzales, and Christie (2006). *Journal of Experimental Social Psychology*, *43*(3), 393-398.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*(1), 1-13.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*, 277-293

R Core Team (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raykov, T. (2001). Bias of coefficient α for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, *25*, 69–76.

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy*, *35*, 299-331.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36.

Schmitz, F., Teige-Mocigemba, S., Voss, A., & Klauer, K. C. (2013). When scoring algorithms matter: Effects of working memory load on different IAT scores. *British Journal of Social Psychology*, *52*(1), 103-121.

Schmukle, S. C., & Egloff, B. (2006). Assessing anxiety with extrinsic Simon tasks. *Experimental Psychology*, *53*(2), 149-160.

Sijtsma, K. (2009). Reliability beyond theory and into practice. *Psychometrika*, *74*(1), 169-173.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 0956797611417632.

Stanley, D., & Phelps, E., & Banaji, M. R. (2008). The Neutral Basis of Implicit Attitudes. *Current Directions in Psychological Science, 17*(2), 164-70.

Uhlmann, E. L., Pizarro, D. A., & Bloom, P. (2008). Varieties of social cognition. *Journal for the Theory of Social Behaviour, 38*(3), 293-322.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, 0734282911406668.

R-code

```
# libraries
require(lavaan)

# Cronbach alpha formula
alpha <- function(data){
    S <- cov(data)
    p <- ncol(S)
    p/(p-1) * (sum(S) - sum(diag(S)))/sum(S)
}

# Read Data (N=322)
# S1 - S2 = Split Half Parcel 1 (2): odd (even) trials
# A1 - A4 = Sequential Parcel 1 (4)
# WP, BP, NP, PP  = White, Black, Negative, Positive Parcel

Data                                                    <-
read.table("http://users.ugent.be/~mldschry/omega/IAT_Data
_Sample_322.dat",               header = TRUE)

# Split-half reliability:
rsh <- cor(Data$S1,Data$S2)
rsh

# Spearman-Brown correction
2*(rsh) / (1+rsh)

# Cronbach alpha
alpha(Data[,c(7:10)])

# 2 Omega using sem
# Standard errors based on bootstrap


# Congeneric Model
M1 <-   '
    L =~ lam1*WP + lam2*BP + lam3*NP + lam4*PP

    WP~~e1*WP
    BP~~e2*BP
    NP~~e3*NP
    PP~~e4*PP

    #Reliability omega
    omega := (lam1 + lam2 + lam3 + lam4)^2 /
             ((lam1 + lam2+ lam3 + lam4)^2 + (e1 + e2 +
     e3 + e4))
```

```
    '

fit1 <- cfa(M1, data=Data, std.lv=TRUE, se="boot")
summary(fit1, fit.measures=TRUE, standardized=TRUE)

# Tau-equivalent Model
M2 <-   '
    L =~ lam1*WP + lam1*BP + lam1*NP + lam1*PP
    WP~~e1*WP
    BP~~e2*BP
    NP~~e3*NP
    PP~~e4*PP
    '

fit2<- cfa(M2, data=Data, std.lv=TRUE)

summary(fit2, fit.measures=TRUE, standardized=TRUE)

# Compare Tau-equivalent Model with Congeneric Model
anova(fit1,fit2)
```

Chapter 6

Unreliable Yet Still Replicable: A Comment on LeBel and Paunonen (2011)[1]

Maarten De Schryver, Sean Hughes, Yves Rosseel, and Jan De Houwer

Abstract

LeBel and Paunonen (2011) highlight that despite their importance and popularity in both theoretical and applied research, many implicit measures continue to be plagued by a persistent and troublesome issue – low reliability. In their paper, they offer a conceptual analysis of the relationship between reliability, power and replicability, and then provide a series of recommendations for researchers interested in using implicit measures in an experimental setting. At the core of their account is the idea that reliability can be equated with statistical power, such that "lower levels of reliability are associated with decreasing probabilities of detecting a statistically significant effect, given one exists in the population" (p.573). They also take the additional step of equating reliability and replicability. In our commentary, we draw attention to the fact that there is no direct, fixed or one-to-one relation between reliability and power or replicability. More specifically, we argue that when adopting an experimental (rather than a correlational) approach, researchers strive to minimize inter-individual variation, which has a direct impact on sample based reliability estimates. We evaluate the strengths and weaknesses of the LeBel and Paunonen's recommendations and refine them where appropriate.

[1] Based on: De Schryver, M., Hughes, S, Rosseel, Y. and De Houwer, J. (2016). Unreliable Yet Still Replicable: A Comment on LeBel and Paunonen (2011). *Frontiers in psychology*, 6:2039. doi: 10.3389/fpsyg.2015.02039

6.1 Introduction

In their original paper, LeBel and Paunonen (2011) draw attention to a measurement revolution that has unfolded within social psychology over the past two decades and that has shaped methodological, theoretical and empirical developments outside of its borders. For much of the past century, researchers relied on a set of *direct procedures* such as semantic differential scales, feeling thermometers and questionnaires when assessing people's attitudes, beliefs and personality characteristics. These procedures are often deployed under the assumption that people not only have introspective access, but also the opportunity and motivation to accurately report on their psychological attributes or content. Yet it is well-known that this assumption is often violated in socially-sensitive situations (e.g., evaluations of racial, gender or religious groups), demand prone domains (e.g., job hiring or clinical assessment contexts) or instances where the individual lacks introspective access to the content under investigation (see Gawronski & Payne, 2010, for a book length treatment).

These limitations sparked a methodological revolution centered on the development and refinement of a new class of *indirect procedures*. At their core, indirect procedures seek to measure in a way that (a) circumvents a person's ability to strategically control their behavior as well as (b) captures psychological processes, attributes or content in ways that does not depend on introspective access. A multitude of indirect procedures have now been developed and many have seen widespread application both inside and outside of psychological science, from clinical psychology (Roefs et al., 2011), to cognitive (Hahn & Gawronski, in press), and developmental psychology (Dunham, Baron, & Banaji, 2008), as well as in neuroscience (Stanley, Phelps, & Banaji, 2008), political (Nosek, Graham, & Hawkins, 2010) and consumer science (Gregg & Klymowsky, 2013). The most influential of these procedures include the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), evaluative priming (e.g., Fazio, 2001) and the Affective Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005) (for more see Gawronski & De Houwer, 2014; Nosek, Hawkins, & Frazier, 2011)[2].

[2] In-line with De Houwer and Moors (2010) we define a procedure as "direct" or "indirect" based on the way in which the measurement context is arranged to capture the behavior of interest (e.g., verbal, speeded categorization of stimuli). We also define the outcome derived from direct procedures as an "explicit" measure, and the outcome derived from indirect procedures as an "implicit" measure based

LeBel and Paunonen (2011) highlight that despite their theoretical and applied implications, the vast majority of implicit measures suffer from unacceptably low levels of reliability, especially when compared to their explicit counterparts (see also Cunningham, Preacher, & Banaji, 2001; Fazio & Olson, 2003; Gawronski et al., 2007; Uhlmann, Pizarro, & Bloom, 2008). These reliability estimates (usually based on split-half correlations or coefficient alphas) range from "abysmally low (Bosson et al., 2000) to moderate (Kawakami & Dovidio, 2001)" (Lebel & Paunonen, p.572) and are argued to have serious knock-on effects for cumulative scientific progress. In their paper, Lebel and Paunonen equate the issue of reliability with the issue of statistical power, and suggest that "lower levels of reliability are associated with decreasing probabilities of detecting a statistically significant effect, given one exists in the population" (p.573). They also take an additional step and equate the issue of reliability with replicability. In particular, they suggest that "random measurement error, which contributes to the unreliability of measures, can prevent an experiment from being exactly repeatable" (p.571). In other words, higher amounts of random measurement error contaminate a measure's score and decreases the likelihood that researchers will be able to replicate their own or other's findings. To put it differently, "given that the probability of replication is simply a special case of statistical power (i.e., probability of replication is the probability of detecting a statistically significant effect given one exists in the population and that the effect has already been found in at least one sample), it follows that decreasing levels of reliability should be associated with reduced likelihood of replication" (p.573).

To test this idea, LeBel and Paunonen (2011) conducted a Monte Carlo simulation to examine the effect of different levels of reliability on the replicability of experimental findings in the context of implicit measures. The authors found that the probability of replicating an experimental effect "systematically decreased as the random measurement error contaminating the scores increased. This pattern was especially pronounced for "medium" and "large" population effect sizes and for moderate to large sample sizes (i.e., N equal to or greater than 30 per condition)" (p.577). Based on the results of their simulation, LeBel and Paunonen put forward three main ideas. First, they argue that random measurement error should be equated with the concept of reliability - and as a result - the probability of replicating an

on the properties of the psychological attribute under investigation. Put simply, "implicit" and "explicit" refer to the operating conditions under which a psychological attribute influences measurement outcomes rather than the procedure itself.

experimental effect decreases as random measurement error (i.e., low reliability) increases. In other words, empirical results that are influenced by random measurement error cannot be replicated exactly whereas results uncontaminated by random measurement error are more likely to replicable (i.e., *probability of replication* increases as a function of *reliability*). Second, they argue that researchers should strive to improve implicit measures that suffer from unacceptable levels of reliability and gravitate towards measures known to have acceptable psychometric properties. Finally, when using implicit measures, researchers should routinely and accurately report reliability, and in the case of experimental work, provide separate reliability estimates for each and every experimental condition.

The above conceptual analysis and associated recommendations certainly seem reasonable on first glance. Yet we believe that these recommendations and the assumptions they are built upon are not as straightforward as one would initially suspect. As we shall see, there is no direct or one-to-one mathematical relationship between the reliability of an implicit measure and the likelihood of replicating an experimental outcome. Random measurement error and reliability refer to two very different psychometric concepts that cannot be used interchangeably. By equating these two concepts, LeBel and Paunonen (2011) arrive at a number of conclusions that might undermine the interpretation and evaluation of data as well as the development of new procedures.

The current commentary has two main goals. First, it aims to provide a quick primer for those interested in the concept of reliability and its relation to implicit measures in experimental contexts. We recognize that this primer will likely contain statistical and psychometric concepts (reliability, power and replicability) that some readers are already familiar with. Our aim is to demonstrate when these concepts are *combined*, a number of conclusions emerge that are, at first sight, counter-intuitive, especially for researchers who are less familiar with psychometric theory and who merely employ implicit measures as tools in their experimental work. Second, LeBel and Paunonen made several recommendations for the experimental use of implicit measures. Like any recommendations, these have the potential to influence the actions of editors and reviewers, as well as the activities of the researcher. We therefore aim to evaluate the strengths and weaknesses of these recommendations, and refine them where appropriate.

6.2 The Relationship between Reliability and Replicability

At the core of LeBel and Paunonen's paper is the notion that reliability is intimately connected with the concepts of statistical power and replicability. To support this assertion, they point to a number of publications demonstrating a positive relationship between the reliability of a dependent variable and the statistical power needed to observe differences between experimental groups or conditions where such differences exist (Rogers & Hopkins, 1988; Sutcliffe, 1958). Yet contrary to their suggestions, the relationship between reliability and statistical power is not a simple, positive or direct one (see Fleiss, 1976; Hopkins & Hopkins, 1979; Overall & Woodward, 1975, 1976; Nicewander & Price, 1978; Overall & Ashby, 1991; Williams & Zimmerman, 1981; Williams, Zimmerman & Zumbo, 1995). For nearly fifty years, the link between reliability and power has been debated in the psychometric literature, with several authors suggesting a positive relation between these two concepts (e.g., Rogers & Hopkins, 1988; Sutcliffe, 1958) while others argue for the very opposite (negative) relationship (e.g., Overall & Woodward, 1975, 1976; Nicewander & Price, 1978). Thus, despite suggestions to the contrary, there appears to be a paradox in arguing for a general or fixed mathematical relation between reliability and power (for more see Williams et al., 1995).

This has serious implications for LeBel and Paunonen's (2011) original argument. If there is no fixed relationship between reliability and power, and if replicability is "simply a special case of statistical power" (p.573), then it follows that there is no general or fixed relation between reliability and replicability. A simple demonstration might help to illustrate our point more clearly. In their original paper, LeBel and Paunonen ran a Monte Carlo simulation to examine the impact of unreliability in a dependent variable on the replicability of results for a simple two-group between-subjects test of means. This simulation revealed that the probability of replicating an experimental effect systematically decreased as the random measurement error contaminating the scores increased. We set out to replicate these findings, but instead of using simulations, we arrived at an exact solution via the formula for calculating power for a two-sample t-test with equal variances (i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$). Working through this example will illustrate the paradox of equating reliability with power or replicability.

First, let X denote observed scores, which can be defined as the sum of unobserved true-scores (T) and error-scores (E). Now, following classical test theory, we can define reliability as the ratio of true-score variance to observed-score variance,

$$\rho_{XX'} = \sigma_T^2/\sigma_X^2, \text{ with } X = T + E,$$

or, we can define reliability in terms of true- and error-score variances,

$$\rho_{XX'} = 1 - \frac{\sigma_E^2}{\sigma_T^2 + \sigma_E^2}$$

and $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$ (Lord & Novick, 1968). Let N be the number of observation in each condition, δ the smallest relevant difference or effect size and $\delta > 0$. Then for a given alpha (α), the power $\pi(\delta)$ can be calculated as follows:

$$\pi(\delta) = 1 - F_{N-1,\frac{\sqrt{N}\delta}{\sigma}}(t_{N-1,\alpha}),$$

where F is the cumulative distribution function of the noncentral t-distribution, with $N - 1$ degrees of freedom and with noncentrality parameter $\sqrt{N}\,\delta/\sigma$.

In their original Monte Carlo simulation, LeBel and Paunonen (2011) fixed the true-score variance (σ_T^2) at 1.00 while allowing the error-score variance (σ_E^2) to vary in order to guarantee *a priori* levels of reliability (i.e., $\sigma_E^2 = (1 - \rho_{XX}) / \rho_{XX}$). Consequently, the observed score variance (σ^2) used in the above power function can be expressed as ($\sigma^2 = \sigma_X^2 = 1.00 + (1 - \rho_{XX}) / \rho_{XX}$). The pattern of results obtained from our power formula for $\rho_{XX} \in \{.10, .20, \ldots, 1.00\}$, $N \in \{10, 20, \ldots, 50\}$, $\alpha = .05$ and $\delta = .50$, can be observed in Figure 1 along-side the original findings from LeBel and Paunonen's (2011) simulation. When true-score variance is fixed, our power function reveals an almost identical (positive) relation between power and reliability as seen in the author's original paper.

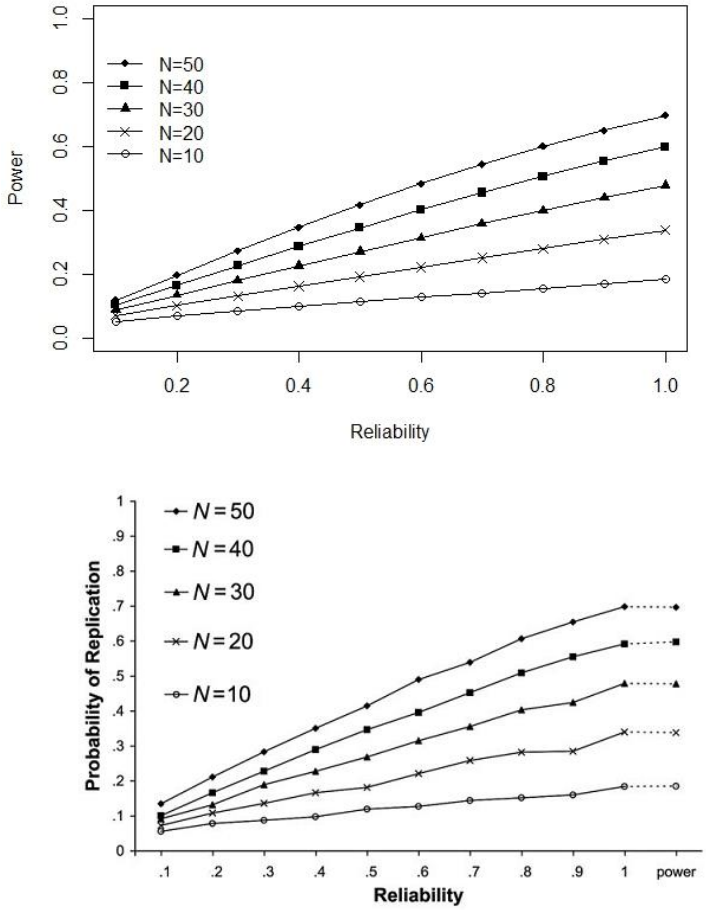


Figure 1. On the upper panel is our exact power estimate as a function of reliability when true-score variance is fixed at 1.00. On the lower panel is LeBel and Paunonen's (2011) simulation of the observed probability of replicating a two-group mean difference effect as a function of sample size (N) and dependent variable reliability (where population effect size equals .5).

Now imagine that instead of true-score variance we fix error-score variance (σ_E^2) at 1.00 and allow the true-score variance to vary as a function of different levels of reliability. In this case the true-score variance as a function of reliability is ($\sigma_T^2 = \frac{1.00}{1-\rho_{XX}} - 1.00$). The observed score variance can then be expressed as ($\sigma^2 = \sigma_X^2 = \frac{1.00}{1-\rho_{XX}}$).

The pattern of results obtained from our power formula for ($\rho_{XX} \in \{.00, .10, \ldots, .90\}$, $\in \{10,20, \ldots ,50\}$, $\alpha = .05$ and $\delta = .50$), can be observed in Figure 2. When error-score variance is fixed, our power function reveals an entirely opposite (negative) relationship between power and reliability as compared to that reported by LeBel and Paunonen (2011).

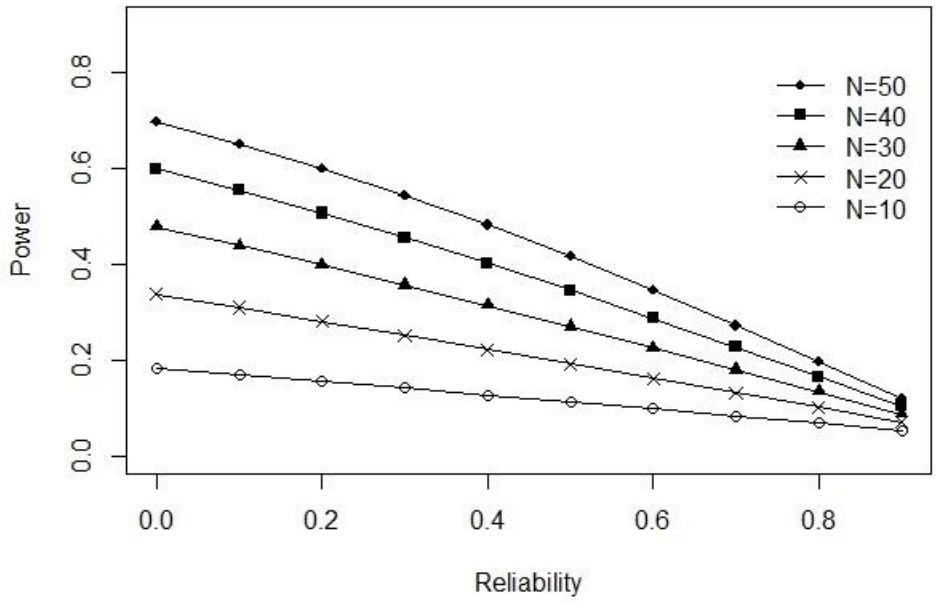


Figure 2. Exact power as a function of reliability when error-score variance is fixed at 1.00.

The above example clearly illustrates the paradox of equating reliability with power or replicability. Consequently, LeBel and Paunonen (2011) do not provide sufficient information to disentangle the various relationships that can potentially exist between reliability and replicability in their original paper. Instead they simply focus on the first of these possibilities (i.e., fixed true-score variance) and thus their conclusions should not be overgeneralized and only applied to such situations.

6.3 Evaluating Research Findings Characterized By Low Levels of Reliability

If it is the case that there is no fixed mathematical relation between reliability and power, then LeBel and Paunonen's second recommendation also needs to be re-examined (i.e., that researchers should "improve those implicit measures having unacceptable levels of reliability

or then utilize implicit measures known to have acceptable psychometric properties"). To illustrate this more clearly, imagine that you are a social psychologist interested in understanding how humans come to like and dislike novel stimuli. You begin by formulating a relatively simple hypothesis that evaluative responses to stimuli can be changed by providing people with verbal information about that stimulus. To test this hypothesis, you provide a group of thirty participants with a set of attitude-relevant instructions (e.g., "*Luupites are good and Niffites are bad*") and another group of thirty participants with attitude-irrelevant instructions (e.g., the basic steps required to waltz at a party). Thereafter, you administer a test of automatic evaluative responding such as an IAT wherein participants have to categorize items related to Luupites and positive words using one response key and items related to Niffites and negative words using another response key. In a second block of trials these response assignments are reversed so that Luupite-related items and negative words are assigned to the first key while Niffite-related items and positive words are assigned to the second key. The difference in performance during the first relative to the second phase (known as the IAT effect) is considered to provide an overall measure of how readily people prefer Luupites compared to Niffites (see De Houwer, 2006, Gregg, Seibt, & Banaji, 2006, for studies along these lines).

Now imagine that data collection is finished. You create a scatterplot and regression line using the IAT scores obtained from the test trials and practice trials for participants in the two instruction conditions (see Figure 3). Analyses reveal that participants provided with attitude-irrelevant instructions displayed a non-significant preference for Niffites over Luupites (M = -0.25, SD = 0.55) while participants provided with attitude-relevant instructions display a clear evaluative bias for Luupites over Niffites (M = 0.68, SD = 0.22). Running a t-test with a Welch's correction reveals a significant difference between the mean preferences of the two experimental conditions, $t(38.11) = 8.54$, $p < .001$.

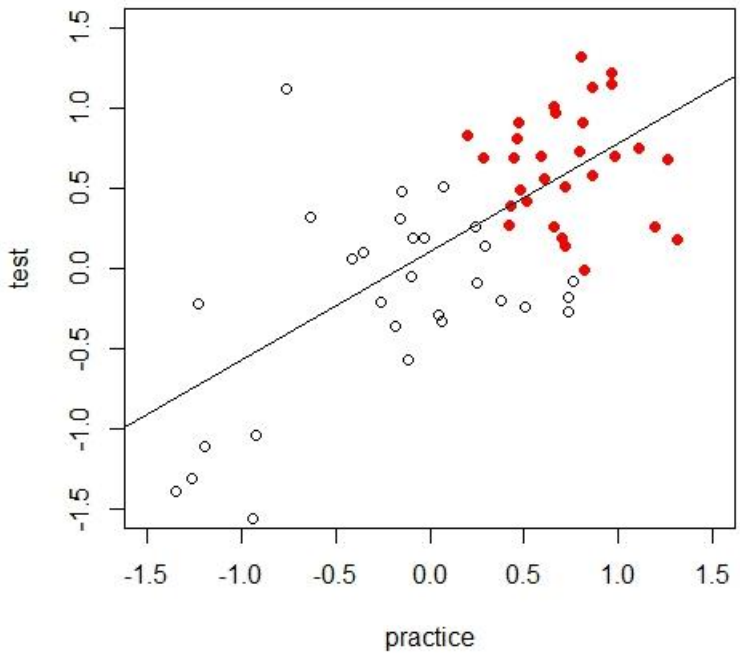


Figure 3. IAT-scores for the practice and test trials for attitude-relevant (filled circles) and attitude-irrelevant instruction (white circles) conditions in our hypothetical example. Note that both practice blocks and test blocks are taken into account for estimating the final IAT-score.

In-line with LeBel and Paunonen's (2011) recommendations, you then estimate the reliability coefficient for both groups using a bootstrap procedure, wherein 1000 random-splits are drawn from the data. For each random split, you estimate a correlation between one split and another. This yields a final reliability estimate in the form of a Spearman-Brown corrected mean split-half correlation. Somewhat surprisingly, you observe a higher reliability estimate for IAT scores in the attitude-irrelevant group (mean r = .92) compared to those in the attitude-relevant group (mean r = .57). The fact that (a) the scores of these two experimental conditions vary in their reliability estimates and (b) the reliability estimate obtained in the attitude-relevant condition is rather low, may cause you as an experimenter, and the individual reviewing your paper, some concern. But is this concern really justified?

The low reliability estimate observed in the attitude-relevant condition tells us that, in this case, *local* measurement precision (due to range restriction in the observed scores) is relatively poor: the relative ordering of participants in this group would probably change if the test was administered under similar contextual conditions. Put another way, we have a relatively homogenous group with respect to the underlying evaluation and our test is not capable of

capturing individual differences *within* that group. But note that this was not the original aim of our study (for more on this point see below). What is important to appreciate here is that the lower level of reliability in the attitude-relevant compared to irrelevant condition does not necessarily imply a higher level of measurement error: if we estimate the group observed-score variances for the attitude-irrelevant ($\sigma_{X1}^2 = .31$) and attitude-relevant instructions conditions ($\sigma_{X2}^2 = .05$) and input these values into the reliability formula ($\sigma_E^2 = \sigma_X^2 - \rho_{XX'}\sigma_X^2$), then the estimated group error-score variance of the attitude-irrelevant group ($\sigma_{E1}^2 = .31 - .92 * .31 = .025$) appears to be slightly larger than that of the attitude-relevant group ($\sigma_{E2}^2 = .05 - .57 * .05 = .022$). In other words, individual IAT-effects in the attitude-relevant group were estimated with a similar level of precision as in the attitude-irrelevant group. The difference in reliability estimates are therefore heavily influenced by differences in true-score variances.

So is it problematic that we observed a rather low reliability score in the attitude relevant condition? The answer - like many in psychological science - is that it depends. Low reliability scores are problematic only if we were interested in differences between individuals (within a group) rather than between groups. Yet in typical experimental designs, including those that use implicit measures, researchers prefer homogenous groups. That is, they strive to decrease observed score-variances within groups or conditions in order to reduce the impact of individual differences, which usually translates into lower true-score variances (Nicewander & Price, 1978, p.407). Such strategies tend to decrease the residual variance in statistical tests such as t-tests or ANOVAs, and as we discussed previously, this often results in lower reliability estimates whenever error-score variances are held constant (also see Williams et al., 2004). Of course, researchers can always improve their measure by replacing their existing test with a tau-equivalent alternative, that is, a comparable test with similar true- but lower error-scores. Doing so will not only lead to a more reliable test, but, due to fixed true-score variance, a more powerful test (as was the case with LeBel and Paunonen original simulation study) (see Nicewander & Price, 1978). However, by replacing one measure with another in situations where their true-scores do not correlate perfectly, researchers introduce uncertainty about the underlying construct in question. Therefore the claim that "researchers need to improve those implicit measures having unacceptable levels of reliability or utilize implicit measures known to have acceptable psychometric properties" should be interpreted with caution.

What about the fact that the reliability estimate in the attitude relevant condition was lower than that in the irrelevant instructions condition? Although LeBel and Paunonen (2011) argue that "differences in observed scores across groups cannot be meaningfully interpreted" in situations where "reliability is drastically different across conditions" (p.580), we argue that even in such cases groups *can* be meaningfully compared, so long as differences in reliability estimates are primarily due to differences in true-score rather than error-score variance (see DeShon, 2004). Thus, in the current example (where error-scores were similar), applying a t-test using Welch's correction will be robust enough to test hypotheses about meaningful mean group differences even though those groups differed in their respective reliability estimates[3].

In short, LeBel and Paunonen's second recommendation should be interpreted with care. The take home message here is that researchers and reviewers should both be aware that low levels of reliability are not necessarily due to increased levels of error-score variance but can also be due to decreased levels of true-score variance. Likewise, the authors' suggestion that some researchers "have been able to easily replicate effects using certain implicit measures, despite their low reliability" (p.579) might reflect the fact that low reliability is sometimes due to reduced true-score variance rather than increased error-score variance. Therefore should researchers try to increase the reliability of implicit measures? On the one hand, we believe that low reliability is acceptable when it occurs due to a reduction in true-score variance. On the other hand, researchers can always improve their (implicit) measure by reducing error as long as this reduction does not affect the variance that is due to the construct of interest. But only by conducting a thorough analysis of different sources of variance can we disentangle these various possibilities.

[3] For comparing more than two means, a non-parametric tests or Bayesian ANOVA can be applied (see Kruschke, 2014), although it should be noted that ANOVAs are rather robust against heteroscedasticity.

6.5 Should Reliability Estimates Be Reported Separately for Each Experimental Condition?

Finally, we agree with the authors that "evaluating (and reporting) the reliability of scores produced by an implicit measure should be viewed as a mandatory requirement when gauging the robustness of a finding" along with the evaluation of sample size, p-values, and confidence intervals". Yet for the reasons noted above, reporting reliability estimates without also providing at least the mean scores and standard deviations of the samples (which would allow the reader to infer true – and error-score variance[4]) is off little value. Moreover, we do not agree that "reliability estimates should be reported separately for each experimental condition", except for situations where the researcher is interested in individual differences within the sample of a particular condition. We are thus somewhat surprised by the example given by LeBel and Paunonen to motivate their argument (see Figure 5, p.580). The authors describe a hypothetical experiment with a control and treatment group that is not unlike our own example above. It is reasonable to assume that (a) these two groups do not initially differ with respect to the underlying attribute of interest or other task-relevant factors such as demographics, (b) do differ after the intended manipulation and that (c) this difference can be observed in their respective implicit test scores. Based on the reliability index of the entire sample (α = .70) and the scatterplot provided by the authors on p.580, their test seems to be a *reliable* and *valid* measure of the underlying attribute. Surprisingly, however, the authors conclude that this reliability index is "artificially inflated due to group mean differences and is completely erroneous" (p.580). They base this conclusion on the reliability estimates obtained from each of the experimental conditions (both $r < .07$), both of which lack internal consistency.

This interpretation seems problematic. In this and other between-groups experiments, the researcher is not interested in examining individual differences within either the control or treatment group. Rather, they are interested in the extent to which individuals from these two groups differ from one another and often use a summary measure (e.g., mean) to do so. Therefore, it seems a little strange to evaluate the implicit measure based on its capacity to

[4] Note that true-score variance can still be influenced by method specific variance (systematic error). More advanced psychometric models could be used to disentangle content specific variance and method specific variance.

detect individual difference within each of the two groups. The reliability estimate for the entire sample and the scatterplot do indicate that the test is capable of detecting differences in the entire sample. Instead of being "artificial" in nature, those differences appear to be due to the intended manipulation and this is illustrated by the fact that there is only a shift in location between the two observed distributions. In other words, the test is doing precisely what the researchers selected it to do. Even if reliability scores were low within, or differed between experimental conditions, this would not be a problem provided that – as we mentioned above – the difference in those reliability estimates was mainly due to differences in true- rather than error-score variance.

In short, the above example seems to be inconsistent with the authors' recommendations. On the one hand, they suggest that researchers "must rule out factors that can reduce the accuracy of reliability estimates, such as the restriction of range… (p.578)" whenever they want to evaluate the reliability of an implicit measure. On the other hand, they suggest that a reliability estimate be calculated for each experimental condition. But this latter suggestion will likely involve reliability estimates that are calculated from a restricted range of scores – a direct contradiction of what the authors recommend above. As we previously mentioned, experimental research typically involves the creation of homogenous groups. A consequence of this is that the range of scores obtained from those groups will likely be *restricted* and thus are not representative of those that would be obtained from a sample representing the entire population.

6.6 Discussion

As Cronbach (1957) eloquently stated "the job of science is to ask questions of Nature" (p.671), and in psychology, these questions have traditionally been asked and answered in two different ways. On the one hand, the correlational approach strives to maximize inter-individual variation in order to explore the relationship between those differences and the phenomenon of interest (i.e., there is a preference for heterogeneous samples). This may be in the service of explaining or predicting when those differences will lead to one outcome versus another. In such a context, the researcher is often interested in maximizing true-score variance so that the test-scores of different individuals can be meaningfully interpreted. On the other

hand, the experimental approach strives to minimize inter-individual variation in order to explore the impact of a particular manipulation on the group as a whole or sub-samples within that group (i.e., there is a preference for homogenous samples). This is often to test causal hypotheses and to make confident causal assumptions about the relationship between one event and another. In such a context, the researcher is typically interested in minimizing true-score variance within conditions so that tests-scores reflect the impact of the intended manipulation rather than erroneous confounds. Thus depending on the scientist's goals and values, the same (implicit) measure may be characterised as either reliable or unreliable as a function of how that researcher responds to true-score variance. High reliability is typically preferable for the correlator while (ironically) the opposite is true for the experimenter because this will lead to a more powerful test. Paradoxically, LeBel and Paunonen (2011) argue that in order to obtain a more powerful test experimenters should strive to develop and use more reliable (implicit) measures.

Yet the paradox for the experimenter is that high observed reliability sometimes leads to more powerful tests and at other times leads to less powerful tests and this makes any discussion about fixed, direct or one-to-one relations between reliability and power or replicability seemingly problematic. On the one hand, we agree with the authors that when true-score variance is fixed an increase in error-score variance will decrease the reliability of a test – and by implication – the likelihood of replication. However, focusing attention on this situation results in an overly simplified view of how reliability relates to replicability that is fraught with conceptual danger (see Nicewander & Price, 1978; Williams et al., 2004 for related arguments). For instance, our own analyses show that it is possible to increase the power of a statistical test (and by implication the likelihood of replication) by decreasing the reliability of an (implicit) measure (e.g., by using more homogeneous samples). It is also the case that implicit measures characterized by low levels of reliability are not necessarily problematic so long as that reliability is a function of reduced true-score variance. Moreover, if researchers aim to explore the reliability of different experimental conditions and report them separately, then low reliability estimates might very well be expected, and even desired. In this case the reliability estimate for the entire sample is not "artificial" but meaningful insofar as it tells us that the measure is capable of detecting individual differences given the range of the true-scores.

Of course we have largely focused on differences in true-score variances throughout our commentary in order to reinforce our central message. Nevertheless, we fully acknowledge that reliability also depends on the amount of error-score variance and that both correlator and experimenter should strive to minimize the impact of this factor where possible. Perugini, Richetin, and Zogmaister (2010) discuss some useful strategies (e.g., using standardized instructions, presenting stimuli in an identical order across participants) that can reduce error-score variance without affecting true-score variance. Also, more advanced psychometric models could be applied to disentangle content specific variance (i.e. true-score variance) from method specific variance (i.e. systematic error-score variance that might influence the true-score variance). For instance, it is well known that measures inferred from raw reaction times can be confounded by general response speed (Fazio, 1990; Faust, Balota, Spieler & Ferraro,1999). By scaling these measures by units of standard deviations, the reliability and validity of these measures can be increased (Greenwald, Nosek & Banaji, 2003; Mierke & Klauer, 2003). Our point is simply that efforts to control error (both random and systematic) will always be important and impact the reliability of an implicit measure in a positive way. But researchers cannot simply equate the former with the latter as LeBel and Paunonen (2011) suggest. Instead, researchers should be aware that low reliability is not always a problem of random measurement error - and in some instances – might actually reflect tight experimental control.

References

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12(11),* 671.

DeShon, R.P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychological Methods, 3,* 412-423.

Dunham, Y., Baron, A. S., & Banaji, M. R. (2008). The development of implicit intergroupcognition. *Trends in Cognitive Sciences, 12(7),* 248-253.

Faust, M.E., Balota, D.A., Spieler, D.H., and Ferraro, F.R. (1999). Individual differences in information processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin,* 125, 777-799.

Fazio, R.H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick & M.S. Clark (Eds.), *Research methods in personality and social* psychology (pp. 74-97), Newburry Park, CA:Sage.

Fazio, R.H. (2001). On the automatic activation of associated evaluations: An overview, *Cognition and Emotion, 15(2),* 115-141.

Fleiss, J.L. (1976). Comment on Overall and Woodward's asserted paradox concerning the measurement of change. *Psychological Bulletin, 83,* 774-775.

Gregg, A. P., & Klymowsky, J. (2013). The Implicit Association Test in market research: *Potentials and pitfalls. Psychology & Marketing, 30(7),* 588-601.

Greenwald, A.G., McGhee, D.E., & Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The implicit Association Test. *Journal of Personality and Social Psychology, 74,* 1464-1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85(2), 197.

Hahn, A., & Gawronski, B. (in press). Implicit social cognition. In J. D. Wright (Ed.), *Theinternational encyclopedia of the social and behavioral sciences* (2nd edition). Oxford: Elsevier.

Hopkins, K.D., & Hopkins B.R. (1979). The effect of the reliability of the dependent variable on power. *Journal of Special Education, 13,* 463-466.

Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd Edition. Academic Press / Elsevier.

LeBel, E.P., & Paunonen, S.V. (2011). Sexy but often unreliable: the impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin*, *37*, 570-583.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA:Addison-Wesley.

Mierke, J., & Klauer, K. C. (2003). Method-specific variance in the implicit association test. *Journal of Personality and Social Psychology*, 85(6), 1180-1192.

Nicewander, W.A., & Price, J.M. (1978). Dependent variable reliability and the power of significance tests. *Psychological Bulletin, 85*, 405-409.

Nosek, B. A., Graham, J., & Hawkins, C. B. (2010). Implicit Political Cognition. In B. Gawronski & B. K. Payne (Eds.), *Handbook of Implicit Social Cognition* (pp 548-564). New York, NY: Guilford.

Nunnaly, J.C. & Bernstein, I.H. (1994). *Psychometric Theory (3rd ed.).* New York: McGraw-Hill.

Overall, J.E. & Ashby, B. (1991). Baseline corrections in experimental and quasi-experimental clinical trials. *Neuropsychopharmacolgy*, *4*, 273-281.

Overall, J.E. & Woodward, J.A. (1975). Unreliability of difference scores: a paradox for the measurement of change. *Psychological Bulletin, 82*, 85-86.

Overall, J.E. & Woodward, J.A. (1976). Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. *Psychological Bulletin*, *83*, 776-777.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes:affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89(3)*, 277.

Roefs, A., Huijding, J., Smulders, F. T., MacLeod, C. M., de Jong, P. J., Wiers, R. W., & Jansen, A. (2011). Implicit measures of association in psychopathology research. *Psychological Bulletin, 137(1),* 149.

Stanley, D., Phelps, E., & Banaji, M. (2008). The neural basis of implicit attitudes. *Current Directions in Psychological Science, 17(2),* 164-170

Sutcliffe, J.P. (1958). Error of measurement and the sensitivity of a test of significance. *Psychometrika*, *23*, 9-17.

Williams, R.H., & Zimmerman D.W. (1981). Error of measurement and statistical inference: Some anomalies. *The Journal of Experimental Education, 49*, 71-73.

Williams, R.H., & Zimmerman D.W., & Zumbo B.D. (1995). Impact of measurement error on statistical power: review of an old paradox. *Journal of Experimental Education, 63*, 363-370.

Chapter 7

General Discussion

This final chapter starts with a brief overview of the five chapters and a summary of the R-code and the R Shiny web apps, that were created in the context of this dissertation. Next, an extensive critical discussion including theoretical implications of the studies is provided. In doing so, theoretical arguments are elaborated by means of additional examples and/or simulations. Also in the discussion of future directions, additional results are presented which serve as further illustrations of the main arguments put forward in this thesis. The chapter ends with a general conclusion.

7.1 Summaries of the studies

In **Chapter 1,** we introduced the Probabilistic Index Models (PIMs) as promising regression models for the effect size PI or P(Y1 < Y2). While for ordinary regression models the associations between an outcome and a covariate is expressed in terms of mean differences, this association is now defined in terms of a probabilistic index (PI). To facilitate the understanding of the PI, we focused on the interpretation of the PI and elaborated on the properties of the PI as an effect size measure. For instance, we demonstrated that the WMW test, of which the PI can be considered as the associated effect size, for many distributions is substantially superior with respect to power when compared to the t-test, even for small sample sizes. We also drew attention to the PI's property of being invariant under monotone transformation. This property could make the PI very attractive for behavioral sciences. When using PIMs, it suffices to assume a monotone relationship between observables and the latent construct to transfer results based on the observed scores to the latent construct.

In **Chapter 2**, we proposed two alternative scoring algorithms for the Implicit Association Test (IAT) whereby the effect is calculated using the PI effect size measure. Apart from its ease of interpretation, we showed that, in contrast to the D-measure, the PI is relatively robust to the presence of outliers and other distributional assumptions, such as skewness. As such, by using the PI effect size measure, measurement precision should be increased. This was reflected in higher reliability estimates. While more or less similar results were obtained with respect to the correlation with explicit measures, a *better* predictive validity of a political IAT was obtained. Besides these nice psychometric properties, the PI scoring algorithm seems to be an especially good alternative for IATs without a built-in error penalty procedure. We have shown that the impact of the proposed error *correction* defined by the D4-scoring algorithm (aka D600) can increase or decrease the total score, depending on the true score. Because of the ordinal way RTs are handled by the PI scoring algorithms, RTs of errors can be easily replaced by the maximum RT of 10,000ms. Irrespective of someone's true score, the effect will always remain the same: errors will be treated as the slowest trial possible.

In **Chapter 3**, the PI-scoring algorithm for the IRAP was introduced. Using a relative large dataset, we have shown that substantial differences were observed between individual D_{IRAP} scores and PI_{IRAP} scores, while a high correlation between those scores was obtained.

Furthermore, we stressed some advantages of using the PI: 1) compared to other robust measures, the PI has a clear and easy interpretation, 2) it allows researchers to calculate confidence intervals and p-values and 3) it allows the inclusion of additional covariates. Furthermore, we argued that these properties make scoring algorithms using a PI-effect size measure especially suitable for single subject design studies.

In **Chapter 4,** a general framework was provided for approximating reliability. We have emphasized the need to start with conceptual clarity and focused on the interaction between sample, test and context when speaking of the reliability of a test. We have emphasized that these elements could vary or remain constant. By providing a Latent Variable Model (LVM) perspective on reliability, a clear way of estimating consistency, equivalence and stability – three proxies of reliability - was provided. In this framework, we discussed the concept of parceling and how the construction of parcels could have an impact on a reliability approximation. To restrict the researchers' degrees of freedom, guidelines for good scientific practice were provided.

In **Chapter 5**, we evaluated the recommendations of LeBel and Paunonen (2011) dealing with reliability issues when using implicit measures. The central issue in their paper is the notion that reliability, statistical power and replicability are closely connected. Through the predominant focus on varying error score variances, these concepts seem to be connected in a simple, positive and direct relationship. However, by fixing the error scores and mainly focusing on varying true score variances, opposite relations between reliability, power and replicability are obtained. Because of this reliability paradox, some inconsistencies in the LeBel and Paunonen (2011) paper were detected and refined.

In the appendices, we provide additional R-code and screenshots of R shiny web apps that we developed. **Appendix 1** includes an R-markdown file with the R code to reproduce all tables, figures and results of Chapter. **Appendix 2** contains screenshots illustrating the R Shiny web app for analyzing IAT data and data obtained from a relational responding task (RRT; De Houwer, Heider, Spruyt, Roets, & Hughes, 2015). This Shiny App allows experimenters to calculate (and download) IAT-effects by 12 different scoring algorithms: the six D-scores algorithms and four conventional algorithms (Greenwald et al., 2003), the G-scoring algorithm (Sriram, Nosek & Greenwald, 2007) and the PI-scoring algorithm (Chapter 2). Furthermore, the program allows for the calculation of the split-half reliability (based on

an odd/even split and corrected by the Spearman-Brown formula), a bootstrap procedure, which involves first splitting the data into two equal halves (random split), calculating a correlation coefficient, and then repeating this process 1000 times (e.g., Ravenzwaaij, van der Maas, & Wagenmakers, 2011). Also, the program enables the user to fit a CFA-model, with one latent variable. In the (lavaan) output, an estimate of coefficient omega is obtained. In **Appendix 3**, a screenshot of the R Shiny web app for IRAP data is provided. This app allows one to calculate for each trial type the PI-score, a 95% confidence interval, the p-value associated with the test H_a: PI $\neq .05$, the standard error for $\widehat{\text{PI}}$, and an estimate of internal consistency (Cronbach's alpha based on two parcel, odd/even split).

7.2 Critical overview

Statisticians, like artists, have the bad habit of falling in love with their models
(G.E.P. Box)

7.2.1 Using PIMs for CRT measures: Did we build a 'Rube Goldberg Machine' ?

In the first three chapters of this dissertation, we showed that a PIM can be considered as a promising alternative to estimate specific effects for CRT tasks. Even though the proposed PIMs have some demonstrably positive properties, one may argue that these models are just too complex for simple IAT and IRAP effects. Hence, scoring algorithms based on PIMs could be considered as some kind of *Rube Goldberg machine*, i.e. a very complex, but fascinating machine developed for one simple task (named after its inventor and cartoonist Rube Goldberg). Perhaps this might reflect the critique of an anonymous reviewer when s/he stated: "*The last thing the IAT literature needs (in my personal view) is yet another complex scoring algorithm.*".

I acknowledge that by using PIMs, the scoring algorithms seem much more complicated and this is potentially in contrast to one of the fundamental principles in science, i.e. "parsimony". In principle, by using Ockham's razor, the discussion of the PI could have been restricted to the two-sample case. As such, the PI could be defined as a scaled Mann-Whitney statistic. Also, when only considering the effect size for the two sample case, our results are *equivalent* with scores obtained from the Gaussian Rank Based or G-scoring algorithm

proposed by Sriram, et al. (2007). Moreover, other authors have proposed several ways of dealing with deviations from normality and methods that are robust to heterogeneity and the presence of outliers (e.g. Richetin, Constantini, Perugini & Schönbrodt, 2015). So why introduce added *complexity*? I believe that the existence of a flexible regression framework that embeds the PI justified an increasing complexity of calculating CRT scores. In what follows, I illustrate some applications of the regression framework.

In Chapter 1, the PI is discussed in detail and we have shown that the effect size is a model-based measure allowing for a more detailed analyses, all within the same framework. For instance, De Schryver (2013) has proposed a PIM to explore the IAT effect for practice and test blocks. The following PIM with a logit link and an interaction term was proposed:

$$\mathrm{P}(Y \leq Y' \mid \boldsymbol{X}, \boldsymbol{X}') = \mathrm{expit}[\beta_1(X_1' - X_1) + \beta_2(X_2' - X_2) + \beta_3(X_1' * X_2' - X_1 * X_2)],$$

where $\boldsymbol{X}^T = (X_1, X_2)$, and X_1 a dummy variable with $X_1 = 0$ for a congruent trial and $X_1 = 1$ for an incongruent trial. X_2 is a dummy variable referring to the phases in the IAT, with $X_1 = 0$ for practice trials and $X_1 = 1$ for test trials.

De Schryver (2013) also proposed a PIM to quantify possible learning effects in the IAT. Therefore, the following PIM was proposed:

$$\mathrm{P}(Y \leq Y' \mid \boldsymbol{X}, \boldsymbol{X}') = \mathrm{expit}[\beta_1(X_1' - X_1) + \beta_2(X_3' - X_3) + \beta_3(X_1' * X_3' - X_1 * X_3)],$$

where X_3 defined a covariate indicating the trial index of each trial. A score of 1 was given to the first six congruent and incongruent trials and a score of 10 was given to the last six congruent trials and incongruent trials (the IAT consisted of 60 congruent and 60 incongruent trials). This model allowed to test the hypothesis of a decreasing IAT-effect as a function of the trial number due to learning during the task: for a randomly selected congruent trial and a randomly selected incongruent trial, both having the same index number, the estimated probability that the congruent trial had a smaller latency equaled $\mathrm{expit}(\hat{\beta}_1 + \hat{\beta}_3 X_3)$.

Another advantage of using PIMs, is that it allows for the construction of confidence intervals for the PI in addition to hypothesis testing. For instance, by fitting the proposed interaction model with two dummy variables (congruent/incongruent trials and practice/test trials) to the data of each participant, the proportion of rejected null hypothesis could be estimated.

7.2.2 R Shiny web Apps, the new SPSS?

In the general introduction, it was argued that new or advanced psychometric and statistical methods are only useful if the end-users of implicit measures (or Choice Reaction Time measures) are encouraged to put these methods into their scientific toolbox and eventually start using them. For this, the availability of the R Shiny apps can be very helpful and can certainly be considered as facilitative towards the integration of psychometrics, statistical theory and psychological theory. Also, it should equalize the calculation for all proposed scoring algorithms, irrespective of the complexity of the statistical theory on which the suggested effect size is built on. However, the availability of these apps can also be considered "*cookbookery*", a term used by the great statistician George Box (1976). By referring to this term, Box (1976) warned for a stagnation of science when researchers are too focused on 'routine techniques' because it can distract them from their initial study goals and might lead to ignorance of testing some necessary model assumptions[1]. I believe that these apps are great for illustrative and educational purposes, but they should be used carefully in research.

7.2.3 On the evaluation of scoring algorithms

The best time to plan an experiment is after you have done it. (Ronald Fisher)

In Chapter 2, it was argued that the PI scoring algorithms are useful because, among other things, these algorithms score *good* on the conventional evaluation criteria. Despite some promising results, these criteria were ignored in our discussion of the PI_{IRAP} in Chapter 3. On the one hand, this is because IRAP users (who typically operate within functional psychology rather than cognitive psychology) are perhaps more interested in individual scores compared to researchers using the IAT. On the other hand, I wonder if the existing criteria are conclusive enough to judge the adequacy of scoring algorithms. In what follows, I briefly discuss the conventional criteria. Next, a more detailed discussion on the appropriateness of four of these criteria is provided. This section ends with a proposal of some alternative criteria. I will argue

[1] An example of the consequences of cookbookery is the frequently used and misused principal component analysis and Cronbach's alpha by SPSS users (Borsboom, 2006)

that the choice of effect size should not depend on a comparative validity and/or reliability study but merely on its appropriateness as a statistical tool to summarize (differences) in behavior.

Psychometric criteria

In the highly cited paper of Greenwald, Nosek and Banaji (2003), the authors identified the following criteria to evaluate candidate measures for the IAT: (1) IAT correlations with explicit measure; (2) correlations of IAT with response latency, (3) internal consistency, (4) Sensitivity to known influences, (5), Resistance to undesired influence of order of combined tasks, (6) Resistance to effect of prior experience taking an IAT. Glashouwer, Smulder, de Jong, Roefs and Wiers (2013) adopted these criteria, with the exception of the fifth one. These authors added the additional criterion of predictive validity. Richetin, et al. (2015) evaluated different scoring algorithms solely on psychometric criteria: (1) reliability, (2) convergent validity with direct measures, (3) convergent validity with other indirect measures, and (4) predictive validity. From this it becomes clear that the main focus to evaluate the appropriateness of different scoring algorithms is on validity and reliability measures.

Before I continue, it is important to disentangle the effect size measure used to summarize the observed behavior and the other *parameters* (i.e., do we need to include the first two trials?, how to treat errors?, do we need to calculate different scores for practice and test trials?, …) defined in a scoring algorithm. Although Greenwald et al. (2003), Glashouwer et al. (2013) and Richetin et al. (2015) have treated the effect size measure as a separate parameter in their evaluations, effect sizes were placed on the same level as other parameters. In what follows, I discuss four important criteria: validity, reliability, correlation with general response speed and sensitivity to known influences.

The Validity criterion Greenwald, et al. (2003) argue that the central assumption in their search for a better scoring algorithm is that "higher implicit-explicit correlations for a modified IAT measure can indicate greater construct validity of the modified measure as a measure of association strength" (p. 200). With *association strength* the authors refer to a latent component (shared variance) of both implicit and explicit scores. They further argue that the

relation between implicit and explicit measures is as similar as the relation between height and weight: rulers for height are better rulers if the correlation between the two measures is higher.

Although Greenwald et al. (2003) and Richetin et al. (2015) both consider the validity criterion superior to other criteria, Glashouwer et al. (2013) argue that this property can be questioned based on dual process theories and only report such correlations to make it possible to compare their results with those from Greenwald et al. (2003)[2]. Briefly, according to dual process theories (see for instance Gawronski & Bodenhausen, 2006) scores may correlate because explicit cognition and implicit *associations* often work together. As such, *higher* or *lower* correlations can be interpreted as respectively *more* or *less* evidence of convergent validity. But, in other conditions it might be that cognition and associations do not synchronize and thus that less positive or even negative correlations can be expected. In these cases, *higher* or *lower* correlations can be interpreted as respectively *less* or *more* discriminant validity.

The discussion of whether convergent or discriminant validity should be the focus of analysis reveals an ambiguity of this property. Even if we agreed that explicit and implicit scores should correlate positively because of a shared underlying latent construct, it would remain unclear why such a (absolute) correlation should be as high as possible. And does a perfect correlation between implicit and explicit measures not imply that we do not need implicit measures at all? Thus far and to the best of our knowledge, no argument can be found that motivates 'the higher is better' assumption.

Furthermore, the implicit and explicit correlations can be seen as part of the nomological network defined to explore *construct* validity of a test. Borsboom, Mellenbergh, and van Heerden (2004) stressed the looseness of such a nomological network due to its unrestrictiveness by allowing for an infinite number of variables into the network. Indeed, in defining the nomological network for exploring construct validity, Cronbach and Meehl (1955) do not state for instance that correlations between constructs should be minimal or maximal; the network is described only in terms of higher and lower correlations. Also, Borsboom, et al. (2004) discuss the usefulness of such a network for construct validity. In doing so, they elaborate on the same example used by Greenwald et al. (2003): "It is even more contrived to presume that the validity of a measurement procedure derives, in any sense,

[2] Richetin et al. (2015) do recognize this dualism, however they made what they called a pragmatic choice and considered only convergent validity.

from the relation between the measured attribute and other attributes. Length is not implicitly defined in terms of its relation with weight, and much less is the validity of a meter stick." (p. 1064).

Until now, we have only discussed the inappropriateness of convergent or discriminant validity, while Glashouwer et al. (2015) and Richetin et al. (2015) also pointed to *predictive validity* as an important criterion. From a strict, pragmatic perspective one could come up with arguments to defend this statement ("*we have no idea why, but we can predict criterion Y…*"). However, this type of validity can only be as good as the validity of the criterion variable itself[3] (see for instance Lord and Novick, 1967) and the adequacy of the specific criterion variables employed can often be questioned (e.g., Talaska, Fiske, & Chaiken, 2008; Carlsson & Agerström, 2016; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013; Gawronski & Bodenhausen, 2017).

The Reliability / Internal consistency criterion In addition to validity, reliability is considered to be one of the most important psychometric properties of a test and this might explain why authors have promoted *internal consistency* as a desired property in evaluating scoring algorithms. Although reliability is considered as the upper bound of validity (see Lord & Novick, 1968), Greenwald et al. (2003) and Richetin et al. (2015) consider the reliability criterion as subordinate to their proposed validity criterion. Richetin et al. (2015) pointed out that some IAT scores (i.e. the effect size measure) might capture "reliable' variance caused by other factors than the latent construct of interest. This might increase the reliability estimate and hence "maximizing reliability in spite of validity is not desirable" (p. 6).

Correlation of implicit measures with response latency Several authors have shown a positive correlation between effect sizes and the general response speed of participants. Greenwald, et al. (2003) argued that good scoring algorithms should take this artifact into account. Differences in general cognitive abilities unrelated to the construct of interest (such as task-switching abilities, working memory capacity, …), could indeed have an impact on the general performance on an implicit task. Indeed, this artifact might have some serious implications. A remarkable example can be found in the developmental psychology literature. Based on a decreasing ratio-effect observed in the numerical distance task, the theory of an increasing numerical representational precision was accepted. However, taking general

[3] Note that this also holds for validity indices.

response speed into account, an increasing ratio-effect was found, which jeopardizes the aforementioned theory (for a detailed overview of this example, we refer to Lyons, Nuerk, and Ansari, 2015).

However, Greenwald et al. (2003) go a step further and argue that scoring algorithms *minimizing* "this undesired artifactual correlation" (p. 200) should be preferred. The same position is taken by Glashouwer et al. (2013). Somewhat surprisingly, the undesired artifactual correlation is then equated with minimizing the correlation coefficient between effect and general response speed.

To disentangle both sources of variance (true score variance and variance related to the artifact) seems rather difficult if not impossible without extra cognitive ability measures (Blanton, Jaccard, Gonzales, & Christie, 2006) or filler trials (Fazio, 1990). However, the solution is perhaps more obvious than difficult. The impact of cognitive abilities on the general performance on a task and hence the general response speed, can be considered as a problem of scales. In other words, RT scales are participant dependent. If one wants to compare the impact of two variables, but both are measured on a different scale (e.g.., variable X has a Likert scale varying from 1 to 7 and the units of variable Y vary from 1 to 100), effects can still be compared by standardizing both scales. For instance, by subtracting scores by the mean and dividing by the standard deviation, the variable is transformed into Z-scores. Because the unit of measurement is cancelled out by the formula, we are able to compare the *relative* effect of both variables (see for instance Kelly & Preacher, 2012; Grissom & Kim, 2012). The same logic can now be applied to implicit measures. If we want to compare the relative effect between participants, we have to cancel out the unit of measurement in calculating individual effects. This type of effect size measures is known as *standardized* effect size measures. As such, every measure that can be considered as a standardized effect size measure (e.g., the D-measures, PI-measure), will meet this property or criterion.

Sensitivity to known influences / modal tendencies This property could be classified under the validity criterion and, hence, the exact same reasoning as for validity can be applied in this paragraph. Nevertheless, it is interesting to discuss this *property* in further detail. The main reason why this is considered as a useful criterion is because it is based on prior knowledge (Greenwald et al., 2003). It seems that because IATs are sensitive to implicit attitudes and stereotypes, typical or expected IAT patterns should be reflected in the scores of

the scoring algorithms. As such, good scoring algorithms should reflect known population differences as much as possible. Because different effect sizes may be defined in different scales (e.g., a scale in milliseconds, a scale varying from -2 to +2, from 0 to 1, etc.), Greenwald et al. (2003) and Glashouwer et al. (2013) choose to express the modal tendency using a Cohens' *d* measure.

Perhaps the most obvious reason why this property can be questioned, lies in the argument of Blanton and Jaccard (2006), stating that measures in psychology are arbitrary. Without going into much detail, the point is that an IAT D-score of zero does not per se mean an absence of bias, and thus middle points are arbitrary. In other words, according to these authors, it makes no sense to test if your score differs from zero. If this reasoning holds true for the IAT and more generally for other CRT tasks, testing if a group mean differs from zero does not make any sense[4].

A less obvious reason is that this property causes contradiction in the list of criteria to evaluate scores. Suppose that two effect size measures need to be compared. First, their reliabilities are estimated, with reliability defined as the true score variance divided by the observed score variance. Also, observed variances are just the sum of the true score variance and the error variance. Now, let the error score variances be fixed and let $\sigma_{E1}^2 = \sigma_{E2}^2 = .20$. For the first test score, the true score variance equals $\sigma_{T1}^2 = .80$, while for the second test score $\sigma_{T2}^2 = .60$. The reliabilities are then $\rho_1^2 = .80$ and $\rho_2^2 = .75$. Based on this analysis, it is concluded that score 1 is the preferred score because it has higher reliability compared to the second score. Now, let us assume that for both scores, the non-standardized means equal $M_1 = M_2 = .50$ based on score 1 and score 2. Transforming these effect sizes to the standardized Cohen's *d* (i.e. $d = M/SD$), with M the sample mean and SD the sample standard deviation, we obtain $d_1 = \frac{M_1}{\sqrt{\sigma_{T1}^2 + \sigma_{E1}^2}} = \frac{.50}{1} = .50$ and $d_2 = \frac{M_2}{\sqrt{\sigma_{T2}^2 + \sigma_{E2}^2}} = \frac{.50}{\sqrt{.80}} = .56$. Based on this analysis, it can be concluded that score 2 is the preferred score because of its larger effect size compared to score 1. Indeed, this is the opposite conclusion compared to the one made when comparing reliability estimates.

[4] In fact, the same holds true when comparing group mean differences, whether it concerns the same group or two different groups.

Alternative criterion: appropriateness as a statistical tool

Based on the aforementioned critical analysis from the current psychometric criteria used to evaluate scoring algorithms, I argue that these criteria are not ideal and are even prone to subjectivity. Because we are all concerned with good scientific practice and replicability, it would be good to have some additional criteria available. As a new criterion I propose to evaluate an effect size measure based on the appropriateness as a statistical tool to summarize (differences) in behavior.

In both Chapter 2 and Chapter 3, we argued our proposal to use another effect size measure from the well-known observation that RT distributions are typically skewed to the right and outliers can be present. Also, the inherent skewness of RT data does not make it obvious to decide which data points can be considered as outliers. As pointed out in the highly (and sometimes wrongly) cited paper of Ratcliff (1993), outlier analysis is much more than just determining a cut-off defined by the mean plus some standard deviations. Also, given the three general characteristics of RT-distributions: 1) non-normal, right-skewed distribution can be expected, 2) skewness increases with increasing task difficulty and 3) standard deviations increase with increasing means (Ratcliff, 2002; Wagenmakers & Brown, 2007), heteroscedasticity between congruent and incongruent experimental blocks can be expected. In other words, it is very likely that in CRT measures of the two distributions under study (the comparison of the RT distributions from the congruent and incongruent blocks) are not within the same location-scale family. To stress that these properties are more present than absent, they are considered as the three *laws* of RT-distributions (Wagenmakers & Brown, 2007) and thus cannot be ignored when choosing appropriate effect size measures.

7.3 Direction for future research

7.3.1 Give meaning to test scores

While the criteria discussed in the previous section strongly focus on validation – it is striking that this focus lies mainly on the third fundamental principle of the nomological network defined by Cronbach and Meehl (1955): how does the construct relate to other constructs (cfr. convergent/discriminant validity and/or predictive validity)? These types of validity can – at best – only give meaning to the construct (Borsboom, et al., 2004). To make meaning *meaningful*, we must at least have an idea about (1) how observed behaviors relate to one other and (2) how we can relate the theoretical construct to the observed behavior. These are the first and the second fundamental principles of the nomological network of Cronbach and Meehl (1955). The answers on these first two principles are important and necessary to understand how the latent construct relates to other constructs. Also, when (parcel) scores are defined as indicators for a latent variable, these scores must, at least, be monotonically related with the latent variable. Furthermore, to examine the validity of a test (i.e., does variation in the construct cause variation in the test scores – see Borsboom, et al., 2004), it is necessary to determine how a test is scored prior to any other analysis.

In doing so, some fundamental principles defining the behavior under study should be specified and evaluated. In many CRT tasks, the fundamental principle is that, given a latent score, participants will respond faster/slower in one condition compared to another (the second principle of the nomological network). If a procedure further states that items in a condition are all sampled from the same stimulus domain, i.e. stimulus equivalence (the first principle of the nomological network), one single effect size measure must be sufficient to summarize this behavior.

To illustrate for instance how the choice of effect size gives meaning to the test score, consider the following example. I sample data for 100 participants. For each participant, twenty congruent RTs are randomly sampled from a normal distribution with M=600 and SD=20, and twenty incongruent RTs are sampled from a normal distribution with M=610 and SD=20. For each participant, the highest RT from the incongruent condition was added with

the participants index (i={1,…,100}). Two effect sizes are calculated: a simple mean difference expressed in raw RT (Figure 1a) and the D-effect size measure (Figure 1b).

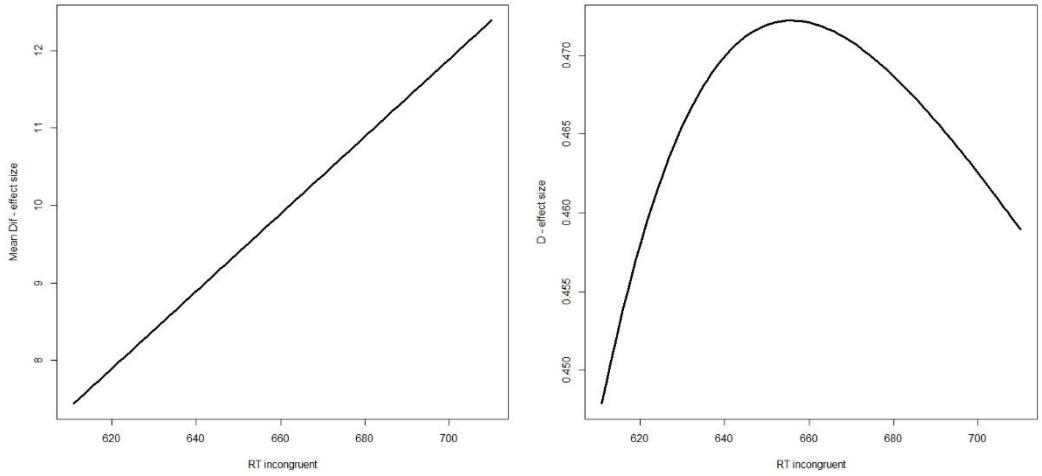


Figure 1. Relationship between mean RT from the incongruent phase and the IAT-effect, calculated as the difference between the mean RT from the congruent and inconguent phase (Figure 1a) or by using the D-effect size measure (Figure 1b).

Whereas a monotonic linear relation between speed and score is observed in Figure 1a, it becomes clear that by using a D-effect size measure a non-monotonic relationship is present. This contradicts the pre-specified laws in the nomological network. Slower incongruent responses are observed for participant 100 compared to participant 50, but a higher test score is observed for the latter compared to the former. From this preliminary observation – and of course more research is needed - it seems that the D-scores should rather be interpreted as the degree of *evidence* that a manipulation (congruent vs incongruent) causes a change in the average behaviors – if the null hypothesis (no changes in the average behavior) is wrong. Although responses are clearly slower for participant 100, the increase of variation in her/his data leads to less evidence.

7.3.2 Invariance under monotone transformations with measurement error

In Chapter 1, we have shown that the PI could be an attractive effect size measure in psychology especially due to its property of being invariant under monotone transformations. This is illustrated by an example in which the relationship between a latent construct (θ, depression) and a proxy measure (Y, BDI-II score) was defined by a non-linear monotone function.

Yet, an important limitation is also pointed out: "these properties only hold under the restrictive setting where we assume a one-to-one mapping between the latent construct and the BDI score". In other words, it is not clear if these properties would still hold in the presence of measurement error. Because measurement error is omnipresent in psychological research and because the aforementioned properties are important for studying psychological latent constructs, we give a preliminary impetus for how to deal with measurement error. In doing so, we will elaborate on the example given in Chapter1.

Let us consider the same non-linear monotone function (*trans()*) to define the relation between θ and Y. Instead of assuming the one-to-one mapping, random measurement error E_i, with $E_i \sim N(0, \sigma_e^2)$ is added. E_i is randomly normally distributed with mean 0, and variance σ_e^2. Our new transformed variable (W) is thus defined by $W = trans(\theta) + E = Y + E$.

From Classical Test Theory (CTT), we define reliability as the ratio of true-score variance to observed score variance, $\rho_{WW'} = \sigma_T^2 \,/\, \sigma_W^2$. Furthermore, $\mathrm{E}(W_i) = \mathrm{E}(Y_i + E_i) = \mathrm{E}(Y_i) + \mathrm{E}(E_i) = \mathrm{E}(Y_i)$. As such, it is shown that Y can be considered as the *true* score of the proxy measure (which is thus not equal to the *true* latent score!). We now can rewrite the reliability formula as follows: $\rho_{WW'} = \sigma_Y^2 \,/\, \sigma_W^2$ and $\sigma_Y^2 = \rho_{WW'}\sigma_W^2$.

In the case of normal distributions[5], we have shown that:

$$\mathrm{P}\,(Y_{IT} < Y_{CT}) = \Phi\left(\frac{\mu_{CT} - \mu_{IT}}{\sqrt{\sigma_{CT}^2 + \sigma_{IT}^2}}\right) (1),$$

[5] If *W* follows a normal distribution, and *W* is defined by its two independent summands *Y* and *E*, then, by Cramér's theorem, *Y* must also be a normal random variable.

and from above we know that $\mu_{CT} = E(W_{CT})$ and $\mu_{IT} = E(W_{IT})$ and $\sigma^2_{CT} = \rho_{WW'}\sigma^2_{W,CT}$ and $\sigma^2_{IT} = \rho_{WW'}\sigma^2_{W,IT}$, under the assumption of measurement invariance. Because P ($T_{IT} < T_{CT}$) = P ($Y_{IT} < Y_{CT}$), the PI corrected for reliability can then be derived by:

$$\text{P}\,(T_{IT} < T_{CT}) = \Phi\left(\frac{E(W_{CT}) - E(W_{IT})}{\sqrt{\rho_{WW'}\sigma^2_{W,CT} + \rho_{WW'}\sigma^2_{W,IT}}}\right) = \Phi(\delta/\sqrt{\rho_{WW'}}).$$

An application of this formula is illustrated in Box 1. We randomly add some error to the (transformed) so that the reliability of the observed variable *W* equals *r* = .60. If no error score variance is taken into account, $\widehat{PI} = .82$, while the true PI equals .88. If we correct for attenuation, we obtain $\widehat{PI} = .88$.

Remember that the derived formula only holds if *W* is normally distributed. For non-normal distributions, the derivation is much more complex and more research is needed.

Box 1. Illustration of the formula correcting the effect size for score reliability.

```
> transf <- function(x, mean = 20, sd = 2){
               qnorm(pt((x-mean)/sd, 30))*sd+mean}
>
> set.seed(2109)
> n <- 10000
> T1 <- rnorm(n, 20, 3)
> T2 <- rnorm(n, 25, 3)
>
> reliability <- 0.60
> varE <-  var(transf(T1))/reliability- var(transf(T1))
>
> W1 <- transf(T1) + rnorm(length(T1),mean = 0, sd = sqrt(varE))
> W1b <-  transf(T1)  + rnorm(length(T1),mean = 0, sd =sqrt(varE))
>
> rel = cor(W1,W1b)
>
> W2 <- transf(T2)+ rnorm(length(T2),mean = 0, sd =sqrt(varE))
>
> wilcox.test(W2,W1)$stat/n^2
0.8169691
> wilcox.test(T2,T1)$stat/n^2
0.8796152
> pnorm((mean(W2) - mean(W1)) /sqrt(var(W1)*rel + var(W2)*rel))
[1] 0.8775691
```

7.3.3 Beyond reliability: stimulus equivalence and measurement invariance

Stimulus equivalence

In the second part of this dissertation, we elaborated on the reliability concept and offered a general framework for estimating reliability measures and their interpretation. Despite the importance of reliability (given that implicit measures typically suffer from low reliability), hitherto little attention had been given to this topic. This is, for instance, reflected in the many different ways in which reliability, mostly operationalized as internal consistency, is estimated in the IAT literature: whereas some authors have used a split-half method, others have used Cronbach's alpha based on *some* (two, three, four, etc.) parcels. Also, our discussion of the reliability 'paradox' as a comment on the paper of LeBel and Paunonen (2011), illustrates that revisiting some properties of reliability can be useful for understanding its relationship with other concepts, such as *power*. This reliability paradox is not only present for implicit measures, but can easily be generalized to all kinds of Choice Reaction Time measures such as, among others, the Stroop task , the Go/No-go task and the Stop-Signal Task (e.g., Hedge, Powell & Summer, 2017).

In Chapter 4, we have empirically illustrated the Latent Variable Model approach for approximating reliability by means of a Race IAT. In this example, we have also demonstrated that reliability estimates are conditional on the person × context interaction. Notwithstanding some differences in reliability are also observed using the Cronbach's alpha formula's (for both the homogeneous parcels data and the distributed parcels data), much more insight is obtained using (a) the Confirmatory Factor Analysis (CFA) approach using (b) homogeneous parcels. Indeed, our results from the Confirmatory Factor Analyses[6] raised several noteworthy issues.

First, a difference in reliability estimates is observed for the White sample ($\hat{r}_W = .80$) compared to the Black sample ($\hat{r}_B = .72$). Moreover, less measurement precision is systematically observed for the homogeneous parcels in the Black sample, with a substantial difference observed for the Black parcel. Second, substantial differences are observed with

[6] Several CFA models were fit to the data from (1) white participants (White sample), (2) black participants (Black sample) and the combined sample. Indicators were defined based on (a) a distributed parceling strategy and (b) a homogeneous parceling strategy, based on trials' respective content: with Black stimuli assigned to a Black parcel, white stimuli to a White parcel, positive adjectives to a Positive parcel, and negative adjectives to a Negative parcel.

respect to the factor loadings across parcels: higher estimates are obtained for the Positive and Negative parcels (i.e., the attributes) compared to the estimates of the Black and White parcels (i.e., the targets). As such, the assumption of exchangeability between items can be questioned. Because the relationship of observed behavior to one other is related the first principle of the nomological network defined by Cronbach and Meehl (1955), the validity of this Race IAT is challenged. Future research should elaborate more on this.

Measurement invariance

The results above touch on a topic seemingly neglected by implicit cognition researchers. From Classical Test Theory, reliability and validity are considered as the two most important criteria for evaluating the psychometric properties of a test. Often, phenomena in two or more distinct groups (or populations) are the subject of the study. For instance, researchers could be interested in implicit attitudes comparing non-addicted and addicted participants. Clinical psychologists could explore and compare implicit self-esteem in both a non-depressed sample and a depressed sample. From our reliability example, using the Race-IAT, we could study effects of in-group/out-group membership (e.g., van Ravenzwaaij, van der Maas, & Wagenmakers, 2010). Likewise, we can compare means within the same population but over different time points. To make these comparisons across groups meaningful, researchers must assure that test scores of the individuals belonging to different groups are on the same measurement scale. This pivotal assumption is known as measurement invariance or measurement equivalence. An extended description of measurement invariance is beyond the scope of this chapter. However, a preliminary illustration might be useful to demonstrate the testable hypotheses relating to measurement equivalence. For an extensive discussion, I refer to Vandenberg and Lance (2000); Meredith and Teresi (2006); and Brown (2014).

Consider the example from Chapter 4. Suppose that researchers want to test the hypothesis that white people and black people differ in implicit in-group/out-group preference. To answer this question, 150 white and 150 black participants are randomly sampled from the Race-IAT of the Attitude 3.0 study from Bar-Anan and Nosek (2014). First, the D1-score is calculated for each participant (Greenwald et al., 2003), with positive scores indicating a stronger association for positive stimuli with the in-group, and for negative stimuli with the out-group. For instance, for a white participant a score of +1 would indicate an implicit

preference for white over black. For a black participant a score of +1 would indicate an implicit preference for black over white. In most research, a simple t-test would be performed to test our hypothesis. The t-test $t(298) = 6.36$, $p < .001$ suggests a difference in mean D1-scores between groups. Based on this analysis, it is concluded that white people have a stronger implicit in-group preference compared to black people. However, before it can be concluded that both groups differ with respect to their implicit in-group preference, one must allow that observed mean score differences can be caused by extraneous elements (Meredith & Teresi, 2006). Hence, measurement invariance should be examined.

Similar to the empirical example from Chapter 4, four homogenous parcels are created: black stimuli are assigned to a Black parcel, white stimuli to a White parcel, positive adjectives to a Positive parcel, and negative adjectives to a Negative parcel. For each participant, four D1-scores are then calculated: one for each parcel. To test for measurement invariance, several hypothesis can be tested. A first test considers a test for configural invariance: does the same congeneric measurement model hold across groups? The second test deals with metric invariance: are corresponding parcel slopes equal? The third test deals with invariant reliability by constraining corresponding parcel slopes to be equal across groups and by constraining the the unique variances to be equal across groups. Comparisons of this model with the less constrained models reveal an inequivalence. Both samples differ with respect to the parcel specific reliability estimates (see Table 2).

Table 2. Parcel reliability estimates for the White sample and Black sample.

	R-WP	**R-BP**	**R-PP**	**R-NP**	**R**
White	.21	.32	.65	.78	.78
Black	.18	.12	.66	.52	.70

So what are the consequences of these inequalities? On the one hand, Raykov can be cited to answer this question (2004, p.310) "[…]reliability inequalities may be associated with even greater discrepancies in validity of construct assessment, thus potentially rendering their group comparisons meaningless.", On the other hand, the impact of the inequality on the parameter of interest can be explored (e.g., Oberski. 2013), and based on this result, it probably can be

concluded that the White and Black sample can be compared anyway. For now, it suffices to stress that the Race-IAT example has no empirical value, but just serves as an illustration in how we can use the reliability framework beyond reliability.

7.4 Conclusion

The introduction of *implicit measures* in the field of social psychology created a new opportunity for RT measures to prove their usefulness as a scientific tool in the discipline of correlational psychology. After all, most implicit measures are based on *choice reaction time* (CRT). Due to the reintroduction of RTs for studying individual differences, experimental and correlational psychology are definitely coming closer again with much more interaction between the two disciplines as a result. Because the two disciplines did not interact that much in the past, each has a particular scientific specialization. By this new interaction, it becomes obvious that both disciplines can learn from each other and as such, psychology as a science can only benefit from this evolution. However, and using Cronbach's (1957) words "It is not enough for each discipline to borrow from the other". Both disciplines can only benefit with a meticulous understanding of each others' scientific goals and methods. In this, psychometric and statistical theory are important for the further development and falsification of psychological tests and theories for both disciplines. With the introduction of Probabilistic Index Models and by offering an extended reliability framework, I hope this dissertation might serve as a catalyst to learning for both disciplines of scientific psychology.

References

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior research methods*, *46*(3), 668-688.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*(1), 27-41.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, *42*(2), 192-212.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425-440.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (**2004**). The Concept of Validity. *Psychological Review*, 111(4), 1061-1071.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791-799.

Brown, T. A. (2014). Confirmatory factor analysis for applied research. Guilford Publications.

Carlsson, R., & Agerström, J. (2016). A closer look at the discrimination outcomes in the IAT literature. *Scandinavian journal of psychology*, *57*(4), 278-287.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American psychologist*, *12*(11), 671-684.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281-302.

De Houwer, J., Heider, N., Spruyt, A., Roets, A., & Hughes, S. (2015). The relational responding task: toward a new implicit measure of beliefs. Frontiers in psychology, 6.

De Schryver, M. (2013). *The Probabilistic Index: A semiparametric scoring framework for the Implicit Association Test.* Unpublished master's thesis, Ghent University, Gent, Belgium.

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. *Research methods in personality and social psychology*, *11*, 74-97.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological bulletin*, *132*(5), 692.

Gawronski, B., & Bodenhausen, G. V. (2017). Beyond Persons and Situations: An Interactionist Approach to Understanding Implicit Bias. *Psychological Inquiry*, 28(4), 268-272.

Glashouwer, K. A., Smulders, F. T., de Jong, P. J., Roefs, A., & Wiers, R. W. (2013). Measuring automatic associations: Validation of algorithms for the Implicit Association Test (IAT) in a laboratory setting. *Journal of Behavior Therapy and Experimental Psychiatry*, *44*(1), 105-113.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of personality and social psychology*, *85*(2), 197.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity. Journal of Personality and Social Psychology, 97(1), 17-41.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications*. Routledge.

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 1-21.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological methods*, *17*(2), 137.

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin*, *37*(4), 570-583.

Lord, F. M., Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Lyons, I. M., Nuerk, H. C., & Ansari, D. (2015). Rethinking the implications of numerical ratio effects for understanding the development of representational precision and numerical processing across formats. *Journal of Experimental Psychology: General*, *144*(5), 1021-1035.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. Medical care, 44(11), S69-S77.

Oberski, D. L. (2013). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. Political Analysis, 22(1), 45-60.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of personality and social psychology*, *105*(2), 171-192.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological bulletin*, *114*(3), 510.

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic bulletin & review*, *9*(2), 278-291.

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. Behavior Therapy, 35(2), 299-331.

Richetin, J., Costantini, G., Perugini, M., & Schönbrodt, F. (2015). Should we stop looking for a better scoring algorithm for handling Implicit Association Test data? Test of the role of errors, extreme latencies treatment, scoring formula, and practice trials on reliability and validity. *PloS one*, *10*(6), e0129601.

Sriram, N., Nosek, B.A., & Greenwald, A.G. (2007). Scale invariant contrasts of response latency distributions. Unpublished manuscript, Univ. Virgina.

Thas, O. (2010). Comparing distributions. Springer.

Talaska, C. A., Fiske, S. T., & Chaiken, S. (2008). Legitimating racial discrimination: Emotions, not beliefs, best predict discrimination in a meta-analysis. *Social justice research*, *21*(3), 263-296.

Wagenmakers, E. J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological review*, 114(3), 830.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, *3*(1), 4-70.

van Ravenzwaaij, D., van der Maas, H. L., & Wagenmakers, E. J. (2011). Does the name-race implicit association test measure racial prejudice? *Experimental psychology*, 58(4), 271-277.

Appendix 1

R tutorial on probabilistic index models

Maarten De Schryver and Jan De Neve

Introduction

This R tutorial is part of the article 'A tutorial on probilistic index models: regression models for the effect size P($Y_1 < Y_2$)' by Maarten De Schryver and Jan De Neve. This document should allow to reproduce all tables, figures and illustrations from the original manuscript. Before you start, it is important you installed the latest R-version (we have used R version 3.4.1). The latest R version can be downloaded from http://cran.r-project.org/.

Two-sample design

Effect measure beyond the mean: weakness and strength

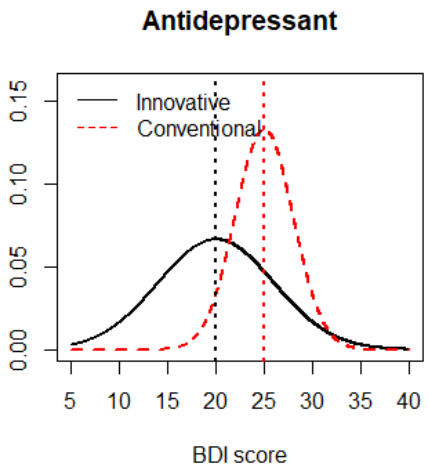
Figure 1

Create artifical data where an innovative treatment outperforms a conventional treatment in terms of BDI scores of patients that receive antidepressants or patients that do not receive antidepressants.

Figure 1a (top left)

The standard deviation is smaller for the BDI scores of the conventional therapy compared to the BDI scores of the innovative therapy.

```
x <- seq(-15, 20, by = 0.01) + 20
plot(x, dnorm(x, 20, 6), type = "l", ylim = c(0, .16), lwd = 2, ylab
= "",
     xlab = "BDI score", main = "Antidepressant")
lines(x, dnorm(x, 25, 3), col = 2, lty = 2, lwd = 2)
abline(v=20, lty = 3, lwd = 2)
abline(v=25, lty = 3, col = 2, lwd = 2)
legend("topleft", c("Innovative", "Conventional"), lty = c(1,2), col
= c(1,2), bty = "n")
```



To calculate the associated PI P($Y_{IT} < Y_{CT}$):

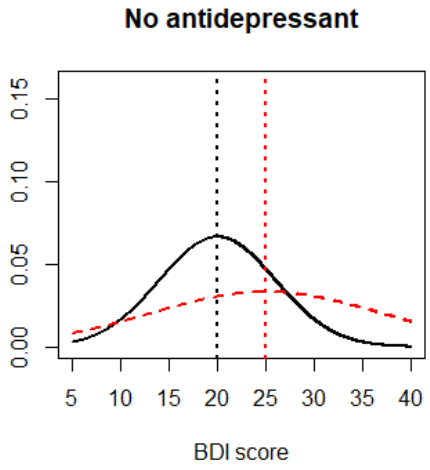
```
pnorm(0, -5, sqrt(36+9))

## [1] 0.7719717
```

Figure 1b (top right)

The standard deviation is larger for the BDI scores of the conventional therapy compared to the BDI scores of the innovative therapy.

```
plot(x, dnorm(x, 20, 6), type = "l", ylim = c(0, .16), lwd = 2, ylab
= "",
     xlab = "BDI score" , main = "No antidepressant")
lines(x, dnorm(x, 25, 12), col = 2, lty = 2, lwd = 2)
abline(v=20, lty = 3, lwd = 2)
abline(v=25, lty = 3, col = 2, lwd = 2)
```



To calculate the associated PI P($Y_{IT} < Y_{CT}$):

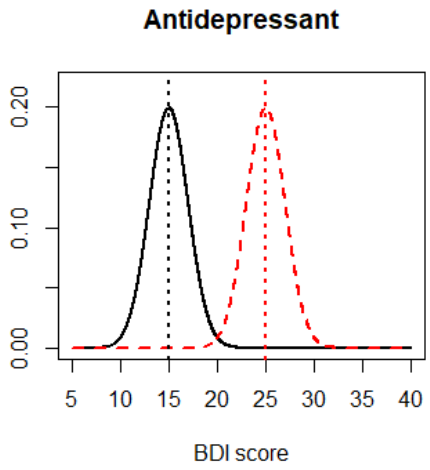
```
pnorm(0, -5, sqrt(36+144))

## [1] 0.6453059
```

Figure 1c (bottom left)

Illustration with non-overlapping densities. The mean difference for the patients receiving antidepressant drugs is 10.

```
x <- seq(-15, 20, by = 0.01) + 20
plot(x, dnorm(x, 15, 2), type = "l", ylim = c(0, .22), lwd = 2, ylab
= "",
     xlab = "BDI score", main = "Antidepressant")
lines(x, dnorm(x, 25, 2), col = 2, lty = 2, lwd = 2)
abline(v=15, lty = 3, lwd = 2)
abline(v=25, lty = 3, col = 2, lwd = 2)
```



To calculate the associated PI P($Y_{IT} < Y_{CT}$):

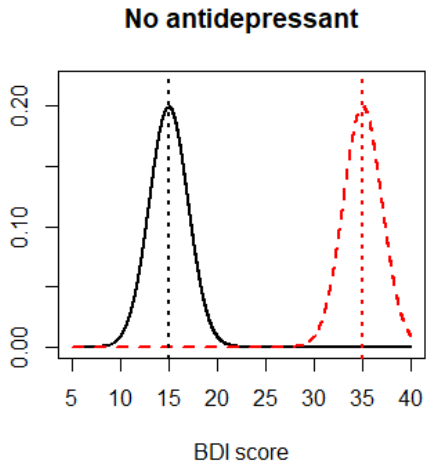
```
pnorm(0, -10, sqrt(4+4))
```

```
## [1] 0.9997965
```

Figure 1d (bottom right)

Illustration with non-overlapping densities. The mean difference for the patients not receiving antidepressant drugs is 20.

```
x <- seq(-15, 20, by = 0.01) + 20
plot(x, dnorm(x, 15, 2), type = "l", ylim = c(0, .22), lwd = 2, ylab
= "",
     xlab = "BDI score", main = "No antidepressant")
lines(x, dnorm(x, 35, 2), col = 2, lty = 2, lwd = 2)
abline(v=15, lty = 3, lwd = 2)
abline(v=35, lty = 3, col = 2, lwd = 2)
```



To calculate the associated PI P($Y_{IT} < Y_{CT}$):

```
pnorm(0, -20, sqrt(4+4))
```

```
## [1] 1
```

Gaining power by exploiting order

For the balanced two-sample designs with 20 or 40 observations per group, power is approximated based on 10,000 Monte-Carlo simulations Because it will take some time to run this code, we have set the number of simulation (N.sim) to 100. To reproduce the results from the manuscript, set N.sim to 10,000.

```
library(rmutil)
```

```
##
## Attaching package: 'rmutil'
```

```
## The following object is masked from 'package:stats':
##
##     nobs
```

```
library(coin)
```

```
## Loading required package: survival
```

```
library(perm)
```

```
N.sim <- 100 # set to 10000 to reproduce our figure in the
manuscript
n.vec <- c(20, 40)
fact.delta <- 0.5
```

```
power.t <- power.wmw <- matrix(nrow = length(n.vec), ncol = 7)
for(j in 1:2){
  n <- n.vec[j]
  for(k in 1:7){
    if(k==1) {err <- rnorm
              delta <- 1}
    if(k==2) {err <- runif
              delta <- 1/sqrt(12)}
    if(k==3) {err <- rlogis
              delta <- pi/sqrt(3)}
    if(k==4) {err <- function(x) rt(x,3)
             delta <- sqrt(3)}
    if(k==5) {err <- rlaplace
              delta <- sqrt(2)}
    if(k==6) {err <- function(x) rt(x,5)
             delta <- sqrt(5/3)}
    if(k==7) {err <- function(x) rexp(x, 1)
              delta <- 1}
    p.t <- p.wmw <- c()
    delta2 <- delta*fact.delta


    for(i in 1:N.sim){
      Y1 <- err(n)
      Y2 <- err(n) + delta2
      X <- factor(c(rep("A",n), rep("B",n)))
      Y <- c(Y1, Y2)
      p.t[i] <- pvalue(oneway_test(Y ~
X,distribution=approximate(B=9999)))
      p.wmw[i] <- wilcox.test(Y1,Y2, exact = TRUE)$p.value
    }
    power.t[j, k] <- mean(p.t < .05)
    power.wmw[j, k] <- mean(p.wmw < .05)
  }
}

colnames(power.wmw) <- c("normal", "uniform", "logistic", "t3",
"laplace", "t5", "exp")
colnames(power.t) <- colnames(power.wmw)
rownames(power.t) <- rownames(power.wmw) <- c("N = 20", "N = 40")

power.t

##        normal uniform logistic   t3 laplace   t5  exp
## N = 20   0.34    0.34     0.38 0.39    0.30 0.37 0.30
## N = 40   0.58    0.59     0.65 0.62    0.62 0.69 0.64

power.wmw
```

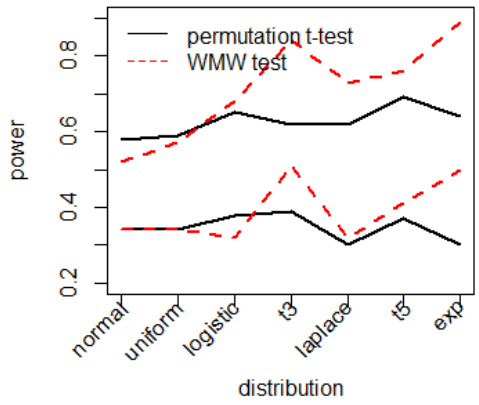
```
##        normal uniform logistic   t3 laplace   t5  exp
## N = 20   0.34    0.34     0.32 0.51    0.32 0.41 0.50
## N = 40   0.52    0.57     0.68 0.84    0.73 0.76 0.89

ARE <- power.wmw/power.t
ARE

##           normal   uniform  logistic       t3  laplace       t5
exp
## N = 20 1.0000000 1.0000000 0.8421053 1.307692 1.066667 1.108108
1.666667
## N = 40 0.8965517 0.9661017 1.0461538 1.354839 1.177419 1.101449
1.390625
```

Figure 2

```
plot(1:7, power.t[1,], ylim = c(0.2,.9), type = "l", lwd = 2,
    xlab= "distribution", ylab = "power", xaxt = "n"  )
lines(1:7, power.wmw[1,], col = 2, lwd = 2, lty = 2)
lines(1:7, power.t[2,], col = 1, lwd = 2)
lines(1:7, power.wmw[2,], col = 2, lwd = 2, lty = 2)
axis(1, at=1:7, labels = FALSE)
text(1:7,par("usr")[3] - .025, srt = 45, adj = 1,
     labels=colnames(power.wmw), xpd=TRUE)
legend("topleft", c("permutation t-test", "WMW test"),
     lty = c(1,2), col = c(1,2), bty = "n")
```



Invariance under monotone transformation

First, we define a monotone transformation function:

```
transf <- function(x, mean = 20, sd = 2){qnorm(pt((x-mean)/sd,
30))*sd+mean}
```

Next, we create artifical latent construct data for 10,000 participants (n) for: the Innovative, Antidepressant condition (T1), the Conventional, Antidepressant condition (T2), the Innovative, No antidepressant condition (T3), the Conventional, Antidepressant condition (T4). The transformation function is then applied to obtain the related observed BDI scores Y1, Y2, Y3 and Y4 and effect sizes are culculated.

```
set.seed(1)
n <- 10000
T1 <- rnorm(n, 20, 3)
T2 <- rnorm(n, 25, 3)
T3 <- rnorm(n, 30, 3)
T4 <- rnorm(n, 35, 3)

Y1 <- transf(T1)
Y2 <- transf(T2)
Y3 <- transf(T3)
Y4 <- transf(T4)

#Calcualte effect sizes:

D1 <- mean(T2) - mean(T1)
D1

## [1] 5.007041

PI1 <- wilcox.test(T2,T1)$stat/n^2
PI1

##         W
## 0.8809579

D2 <- mean(Y2) - mean(Y1)
D2

## [1] 4.557494

PI2 <- wilcox.test(Y2,Y1)$stat/n^2
PI2

##         W
## 0.8809579

D2 - D1

## [1] -0.4495472

PI2 - PI1

## W
## 0
```

```
D3 <- mean(T4) - mean(T3)
D3

## [1] 4.960697

PI3 <- wilcox.test(T4,T3)$stat/n^2
PI3

##         W
## 0.8779221

D4 <- mean(Y4) - mean(Y3)
D4

## [1] 2.746572

PI4 <- wilcox.test(Y4,Y3)$stat/n^2
PI4

##        W
## 0.877922

D4 - D3

## [1] -2.214125

PI4 - PI3

##      W
## -5e-09
```

Figure 3 (middle panel)

```
set.seed(2)
x <- seq(10,45,by=0.1)
plot(x, transf(x, mean = 20, sd = 2), type = "l",
     xlab = "Latent construct", ylab = "Observed BDI score", lwd =
2)
abline(0,1, col = 2, lty = 2, lwd = 2)
```

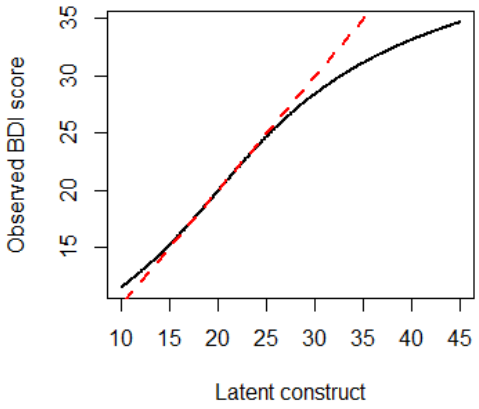


Figure 3 (top left)

```
plot(density(T1, adjust = 2), type = "l", ylim = c(0, .15), xlim =
c(10, 45),lwd = 2, ylab = "",
     xlab = "Latent construct", main = "Antidepressant", yaxt =
"none")
lines(density(T2, adjust = 2), col = 2, lty = 2, lwd = 2)
legend("topleft", c("Innovative", "Conventional"), lty = c(1,2), col
= c(1,2))
```

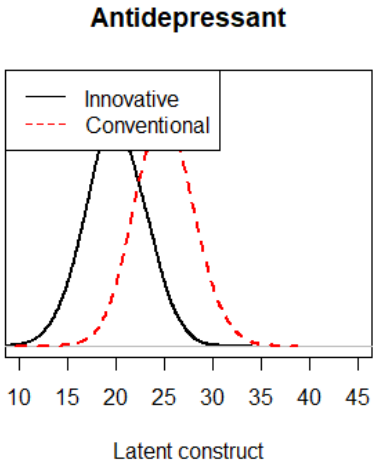


Figure 3 (bottom left)

```
plot(density(Y1, adjust = 2), type = "l", ylim = c(0, .15), lwd = 2,
ylab = "",
     xlab = "Observed BDI score", main = "Antidepressant",xlim =
```

```
c(10, 45), yaxt = "none")
lines(density(Y2, adjust = 2), col = 2, lty = 2, lwd = 2)
```

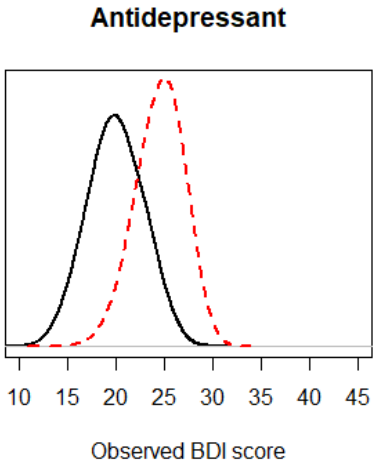


Figure 3 (top right)

```
plot(density(T3, adjust = 2), type = "l", ylim = c(0, .15), xlim =
c(10, 45),lwd = 2, ylab = "",
     xlab = "Latent construct", main = "No antidepressant", yaxt =
"none")
lines(density(T4, adjust = 2), col = 2, lty = 2, lwd = 2)
```

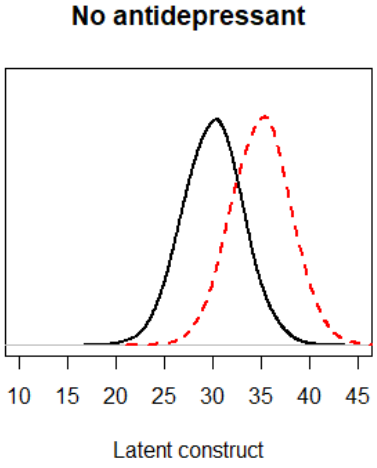
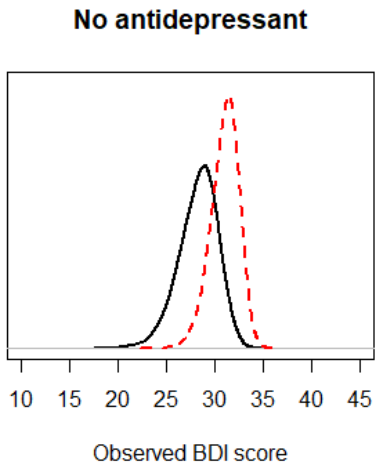


Figure 3 (bottom right)

```
plot(density(Y3, adjust = 2), type = "l", ylim = c(0, .3), lwd = 2,
ylab = "",
     xlab = "Observed BDI score", main = "No antidepressant",xlim =
c(10, 45), yaxt = "none")
lines(density(Y4, adjust = 2), col = 2, lty = 2, lwd = 2)
```



Illustration

BtheB study

We illustrate the methodology on data from a clinical trial to evaluate a computerized, interactive cognitive behavioural therapy for patients with depression. The original study is reported in Proudfoot et al. (2003) and part of the data are available in the R package of Hothorn and Everitt (2017a,b). Patients with depression were recruited in primary care and were randomized over two treatments: an innovative treatment or a conventional treatment. The conventional treatment (TAU: treatment as usual) consisted of face-to-face cognitive behavioral therapy, while the innovative treatment consisted of an interactive computerized program called Beat the Blues (TM) (BtheB) replacing the face-to-face counseling. We refer to Proudfoot et al. (2003) for details.

We consider the following variables: the Beck Depression Inventory II score at baseline (bdi.pre) and after three months (bdi.3m), the treatment (treatment: TAU or BtheB) and weather the patient takes anti-depressant drugs (drug: yes or no). It is of interest to study 1) the association between the treatment and the depression score and 2) the association between antidepressants and the depression score.

```
#load the required packages
library(HSAUR3)

library(DAAG)

library(pim)

library(MASS)

library(geoR)



#assign part of the BtheB-dataset to new object Data - missing
values are removed
Data <- na.omit(BtheB[,c(4,6,3,1)])
rownames(Data) <- NULL
summary(Data)

##     bdi.pre          bdi.3m      treatment   drug
##  Min.   : 7.00   Min.   : 0.00   TAU  :36   No :41
##  1st Qu.:15.00   1st Qu.: 6.00   BtheB:37   Yes:32
##  Median :21.00   Median :13.00
##  Mean   :23.15   Mean   :14.81
##  3rd Qu.:31.00   3rd Qu.:20.00
##  Max.   :47.00   Max.   :53.00
```

Figure 4 (top left)

```
boxplot(bdi.3m~treatment, data = Data, ylab = "Depression score at 3
months",
        xlab = "Treatment", names = c("Conventional", "Innovative"))
stripchart(bdi.3m~treatment, data = Data, vertical=T,
method="jitter",
           add=TRUE, pch=21, col = "gray")
```

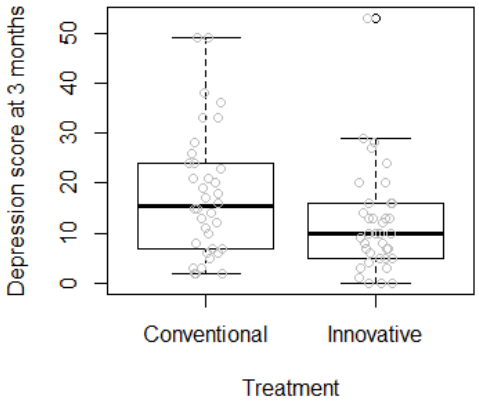


Figure 4 (top right)

```
boxplot(bdi.3m ~ drug, data = Data, ylab = "Depression score at 3
months",
        xlab = "Drugs")
stripchart(bdi.3m ~ drug, data = Data, vertical=T, method="jitter",
           add=TRUE, pch=21, col = "gray")
```

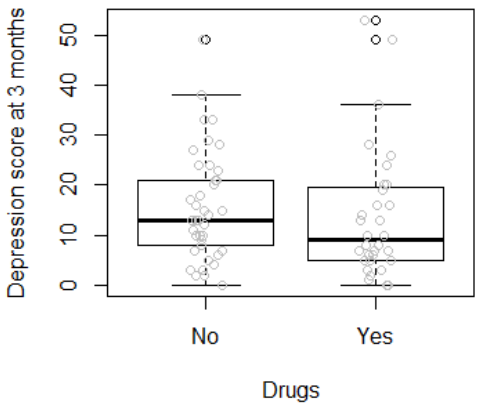


Figure 4 (bottom left)

```
boxplot(bdi.pre ~ drug, data = Data, ylab = "Depression score at
baseline",
        xlab = "Drugs")
stripchart(bdi.pre ~ drug, data = Data, vertical=T, method="jitter",
           add=TRUE, pch=21, col = "gray")
```

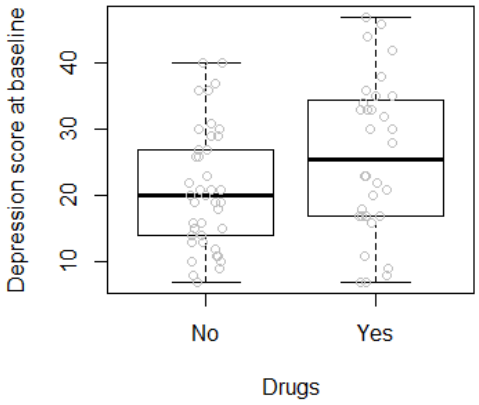
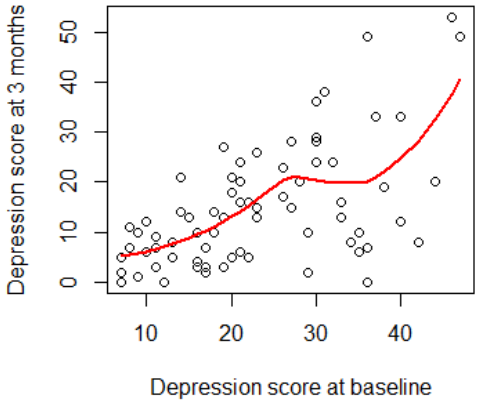


Figure 4 (bottom right)

```
plot(bdi.3m ~ bdi.pre, data = Data, xlab = "Depression score at
baseline",
     ylab = "Depression score at 3 months")
DataSort <- Data[order(Data$bdi.pre),]
lines(DataSort$bdi.pre, predict(loess(bdi.3m ~ bdi.pre, data =
DataSort, span = .75),
                                DataSort$bdi.pre), col = 2, lwd=2)
```



Association between the treatment and the depression score

Two-sample t-test

```
t.test(bdi.3m~treatment, var.equal=TRUE, data=Data)
```

```
##
##  Two Sample t-test
##
## data:  bdi.3m by treatment
## t = 2.0849, df = 71, p-value = 0.04067
## alternative hypothesis: true difference in means is not equal to
0
## 95 percent confidence interval:
##   0.2461034 11.0331759
## sample estimates:
##   mean in group TAU mean in group BtheB
##            17.66667            12.02703
```

Equivalent, via a linear regression model:

```
fit.lm <- lm(bdi.3m~treatment, data=Data)
coef(summary(fit.lm))

##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    17.66667   1.925751  9.173908 1.121148e-13
## treatmentBtheB -5.63964   2.704960 -2.084926 4.067314e-02
```

Repeating the two-sample t-test without outliers (value 53 at row 62)

```
t.test(bdi.3m~treatment, var.equal=TRUE, data=Data[-62,])

##
##  Two Sample t-test
##
## data:  bdi.3m by treatment
## t = 2.7323, df = 70, p-value = 0.007956
## alternative hypothesis: true difference in means is not equal to
0
## 95 percent confidence interval:
##   1.830281 11.725275
## sample estimates:
##   mean in group TAU mean in group BtheB
##            17.66667            10.88889
```

Wilcoxon-Mann-Whitney

```
wilcox.test(bdi.3m~treatment,exact=FALSE, data=Data)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  bdi.3m by treatment
## W = 851.5, p-value = 0.04104
## alternative hypothesis: true location shift is not equal to 0
```

Calculate the PI:

```
n1 <- length(which(Data$treatment=="TAU"))
n2 <- length(which(Data$treatment=="BtheB"))
W <- wilcox.test(bdi.3m~treatment,exact=FALSE, data=Data)$statistic
PI <- W/(n1*n2)
PI

##         W
## 0.6392643
```

Using the R-package pim

```
fit.pim <- pim(bdi.3m~treatment,link="probit", data=Data)
summary(fit.pim)

## pim.summary of following model :
##  bdi.3m ~ treatment
## Type:  difference
## Link:  probit
##
##
##                Estimate Std. Error z value Pr(>|z|)
## treatmentBtheB  -0.3565     0.1703  -2.093   0.0363 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Null hypothesis: b = 0
```

For calculating the estimated probability that a patient receiving the Innovative treatment (BtheB) will have a larger depression score as compared to a patient receiving the Conventional treatment (TAU), we use the pnorm function:

```
pnorm(coef(fit.pim))

## treatmentBtheB
##      0.3607357
```

the 95% confidence interval can easily be obtained by:

```
confint(fit.pim)

##                      2.5 %      97.5 %
## treatmentBtheB -0.6902768 -0.02270898

pnorm(confint(fit.pim))

##                     2.5 %    97.5 %
## treatmentBtheB 0.2450101 0.4909412
```

Logit link

A similar result is obtained when using the `logit' link function. The results are even identical when we transform back to the PI scale.

```
fit.pim.logit <- pim(bdi.3m~treatment,link="logit", data=Data)
summary(fit.pim.logit)

## pim.summary of following model :
##  bdi.3m ~ treatment
## Type:  difference
## Link:  logit
##
##
##                Estimate Std. Error z value Pr(>|z|)
## treatmentBtheB  -0.5722     0.2765   -2.07   0.0385 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Null hypothesis: b = 0

plogis(coef(fit.pim.logit))

## treatmentBtheB
##      0.3607357

plogis(confint(fit.pim.logit))

##                     2.5 %    97.5 %
## treatmentBtheB 0.2471148 0.4924294
```

The association between antidepressants and the depression score

Wilcoxon-Mann-Whitney

```
wilcox.test(bdi.3m~drug,exact=FALSE, data=Data)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  bdi.3m by drug
## W = 772, p-value = 0.1987
## alternative hypothesis: true location shift is not equal to 0
```

Using the R-package pim

```
fit.pim2 <- pim(bdi.3m~drug,link="probit", data=Data)
summary(fit.pim2)

## pim.summary of following model :
##  bdi.3m ~ drug
## Type:  difference
```

```
## Link:  probit
##
##
##         Estimate Std. Error z value Pr(>|z|)
## drugYes  -0.2235     0.1718    -1.3    0.193
##
## Null hypothesis: b = 0
```

```
pnorm(coef(fit.pim2))
```

```
##   drugYes
## 0.4115854
```

```
pnorm(confint(fit.pim2))
```

```
##             2.5 %    97.5 %
## drugYes 0.2876416 0.5451236
```

Depression score at baseline as confounder

```
fit.pim3 <- pim(bdi.3m~drug+bdi.pre,link="probit", data=Data)
```

```
## Warning: nleqslv says: x-values within tolerance 'xtol'
##  See ?nleqslv for more info.
```

Sometimes another estimator will be needed for solving the score function. For instance, we can use the more `stable', but slower estimator.

```
fit.pim3 <- pim(bdi.3m~drug+bdi.pre,link="probit",estim =
estimator.glm, data=Data)
summary(fit.pim3)
```

```
## pim.summary of following model :
##  bdi.3m ~ drug + bdi.pre
## Type:  difference
## Link:  probit
##
##
##          Estimate Std. Error z value Pr(>|z|)
## drugYes -0.522230   0.185775  -2.811  0.00494 **
## bdi.pre  0.049163   0.009856   4.988 6.09e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Null hypothesis: b = 0
```

```
pnorm(coef(fit.pim3))
```

```
##   drugYes   bdi.pre
## 0.3007552 0.5196055
```

```
pnorm(confint(fit.pim3))
```

```
##              2.5 %    97.5 %
## drugYes 0.1877168 0.4371819
## bdi.pre 0.5119053 0.5272983
```

When we account for the baseline depression score, the effect of antidepressant becomes significant. When we compare two patients with the same BDI score at baseline, the probability that the BDI score at 3 months is lower for the partient that does not take antidepressants is estimated at 30%.

The probability that a patient who had a 10 unit higher baseline BDI score compared to another patient with the same antidepressant status can be estimted via:

```
pnorm(10*coef(fit.pim3)[2])
```

```
##   bdi.pre
## 0.6885109
```

Goodness-of-fit

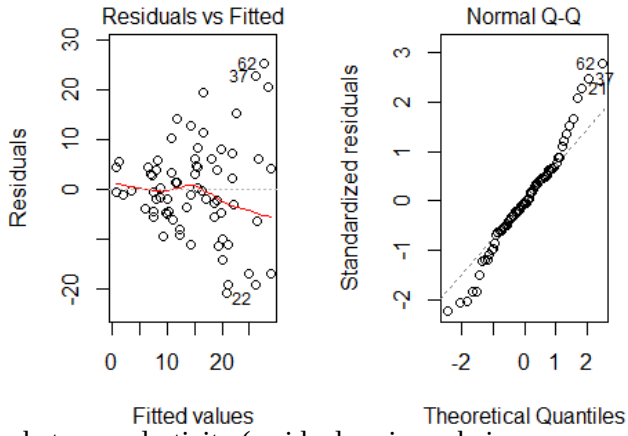
The first step is to fit the probit model. Here, we will explore the consistency of our last pim-model () with the underlying data-generating model.

Step 1: Fit the corresponding linear model

```
fit.lm3 <- lm(bdi.3m~drug+bdi.pre, data=Data)
```

Step 2: Check the assumptions of the linear model

```
par(mfrow=c(1,2))
plot(fit.lm3, which=c(1,2))
```



Because we observe some heteroscedasticity (residual variance being non-constant), we go to step 4.

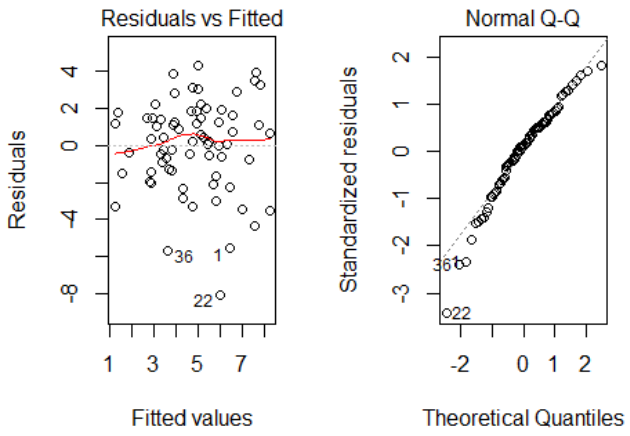
Step 4: Perform a Box-Cox transformation on the linear model.

Because the presence of non-positive values in our data, we opt for the two-parameter Box-Cox tranformation. In contrast to the more common (one-parameter) Box-Cox transformation, a shift parameter (lambda2) is additionaly estimated. In cases this parameter is zero, the model becomes the one-parameter Box-Cox transformation.

```
library(geoR)
bc <- boxcoxfit(Data$bdi.3m, cbind(Data$drug, Data$bdi.pre),
lambda2=TRUE)
bc$lambda

##    lambda   lambda2
## 0.4675885 0.0005300

lambda <- bc$lambda[1]
lambda2 <- bc$lambda[2]
if(lambda==0){trans.bdi.3m <- log(Data$bdi.3m + lambda2)}
if(lambda!=0){trans.bdi.3m <- ((Data$bdi.3m + lambda2) ^ lambda - 1)
/ lambda}
fit.bc<- lm(trans.bdi.3m~drug+bdi.pre, data=Data)
par(mfrow=c(1,2))
plot(fit.bc, which=c(1,2))
```

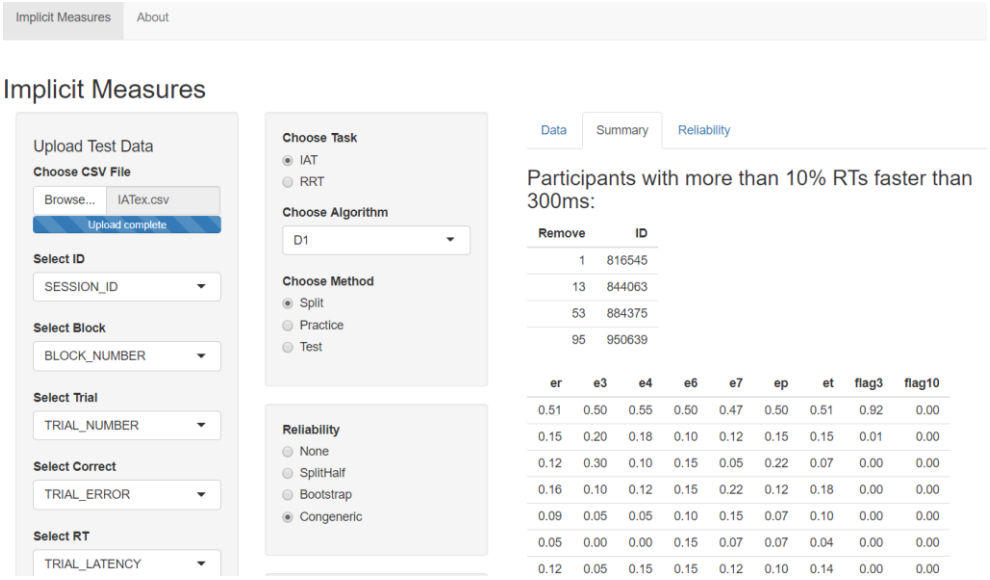


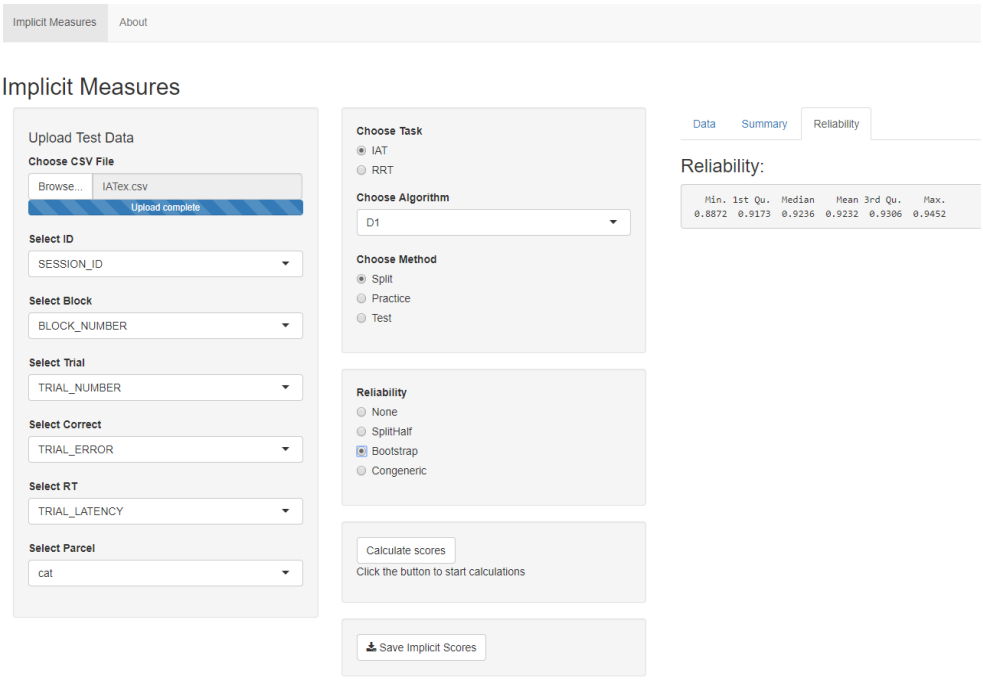
The residual plot suggests that by transforming the data, the variance is stabilized. Also, the residuals are more or less normal. We may go to step 6 and we conclude that the PIM with probit function will be consistent with the underlying data-generating model.

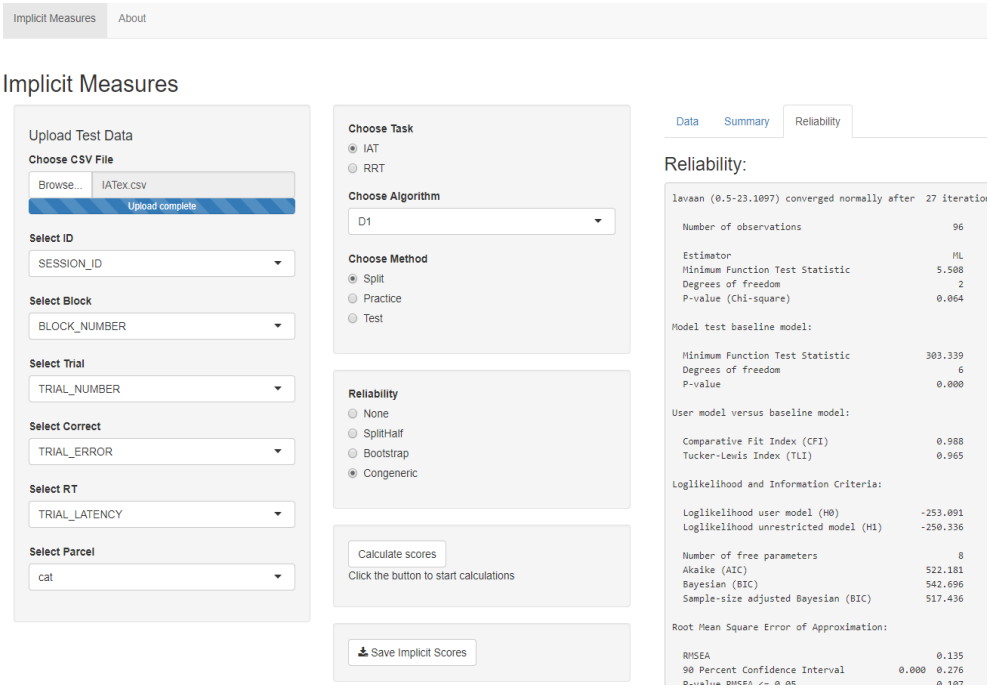
Appendix 2

R Shiny App for IAT and RRT

URL: http://datapp.ugent.be/shiny/implicit/




Implicit Measures
About
Implicit Measures
Upload Test Data
Choose CSV File
Browse...
IATex.csv
Upload complete
Select ID
SESSION_ID
Select Block
BLOCK_NUMBER
Select Trial
TRIAL_NUMBER
Select Correct
TRIAL_ERROR
Select RT
TRIAL_LATENCY
Select Parcel
cat
Choose Task
IAT
RRT
Choose Algorithm
D1
Choose Method
Split
Practice
Test
Reliability
None
SplitHalf
Bootstrap
Congeneric
Calculate scores
Click the button to start calculations
Save Implicit Scores
Data
Summary
Reliability
Reliability:
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.8872 0.9173 0.9236 0.9232 0.9306 0.9452

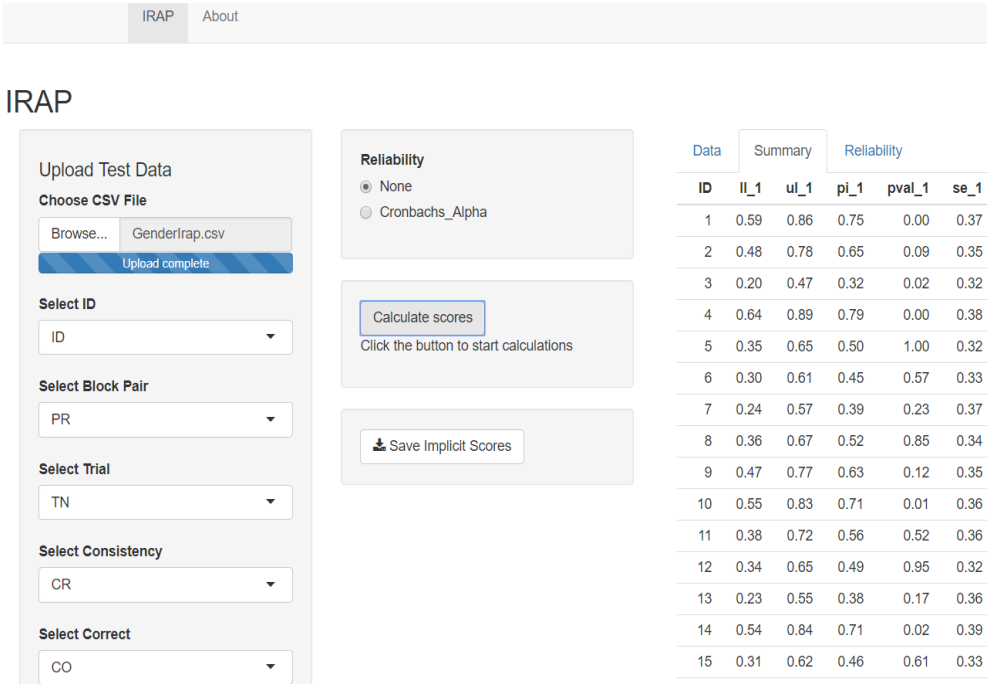

Implicit Measures
About
Implicit Measures
Upload Test Data
Choose CSV File
Browse...
IATex.csv
Upload complete
Select ID
SESSION_ID
Select Block
BLOCK_NUMBER
Select Trial
TRIAL_NUMBER
Select Correct
TRIAL_ERROR
Select RT
TRIAL_LATENCY
Select Parcel
cat
Choose Task
IAT
RRT
Choose Algorithm
D1
Choose Method
Split
Practice
Test
Reliability
None
SplitHalf
Bootstrap
Congeneric
Calculate scores
Click the button to start calculations
Save Implicit Scores
Data
Summary
Reliability
Reliability:
lavaan (0.5-23.1097) converged normally after 27 iteratio
Number of observations 96
Estimator ML
Minimum Function Test Statistic 5.508
Degrees of freedom 2
P-value (Chi-square) 0.064
Model test baseline model:
Minimum Function Test Statistic 303.339
Degrees of freedom 6
P-value 0.000
User model versus baseline model:
Comparative Fit Index (CFI) 0.988
Tucker-Lewis Index (TLI) 0.965
Loglikelihood and Information Criteria:
Loglikelihood user model (H0) -253.091
Loglikelihood unrestricted model (H1) -250.336
Number of free parameters 8
Akaike (AIC) 522.181
Bayesian (BIC) 542.696
Sample-size adjusted Bayesian (BIC) 517.436
Root Mean Square Error of Approximation:
RMSEA 0.135
90 Percent Confidence Interval 0.000 0.276

Appendix 3

R Shiny App for IRAP

Maarten De Schryver and Ian Hussey

URL: http://datapp.ugent.be/shiny/irap/



English Summary

A psychometric analysis of choice reaction time measures

Measures of psychological attributes, such as personality and attitudes, play a central role in the development of psychological theories and their application in daily life. Because many psychological attributes are not directly observable, attribute measures are often based on observable responses to construct-related stimuli in a structured environment. Attributes are therefore often measured through self-report questionnaires. However, explicit self-reports strongly depend on the willingness and ability of respondents to report attributes or behavior. For instance, socially desirable or strategic responding, as well as respondents' unawareness of attitudes are well-known context factors that can bias measures based on self-reports (Wittenbrink & Schwarz, 2007). In view of these concerns, implicit measures were developed, most of which are based on choice reaction time tasks. Well-known examples of such reaction time measures are the evaluative priming task (Fazio, Jackson, Dunton & Williams, 1995), the Implicit Association Test (Greenwald, McGhee & Schwartz, 1998) and the Implicit Relational Assessment Procedure (Barnes-Holmes, Barnes-Holmes, Stewart & Boles, 2010).

Whereas behavior on explicit self-reports is under direct control of participants, implicit measures are assumed to capture uncontrolled and unintentional response tendencies, which are considered as a proxy of a particular attribute. The introduction of choice reaction time measures raised new psychometric challenges, for instance, related to reliability and validity of these tests (De Houwer, Teige-Mocigemba, Spruyt & Moors, 2009; LeBel & Paunonen, 2011), the parsimony of the scoring procedures used to analyze the data they generate (Blanton, Jaccard, Gonzales & Christie, 2006), and the identification of spurious factors that bias these measures (Fiedler, Messner & Bluemke, 2006). Also, the application of choice reaction time measures in (clinical) assessment situations as diagnostic tools strengthens the need for psychometric models of indirect reaction time measures that take into account individual differences.

Within this context, the present PhD project further investigates the psychometric properties of implicit reaction time measures, enhanced them if necessary and develops novel

approaches in testing and enhancing these properties. To this end, insights from psychometric modeling theory and new statistical methods are integrated. **Chapter 1** starts with a general introduction of psychometrics and measurement in psychology. It continues with a reflection on potential challenges of implicit measures and the presentation of the overall objective of this doctoral thesis. In **Chapter 2** the Probabilistic Index Models (PIMs) are introduced as promising regression models for the effect size PI or $P(Y_1 < Y_2)$. In **Chapter 3**, we propose two alternative scoring algorithms for the Implicit Association Test (IAT) whereby the effect is calculated using the PI effect size measure. In **Chapter 4**, the PI-scoring algorithm for the Implicit Relational Assessment Procedure is introduced. In **Chapter 5,** a general framework based on Confirmatory Factor Analysis is provided for approximating reliability. In **Chapter 6**, we evaluate the recommendations of LeBel and Paunonen (2011) dealing with reliability issues when using implicit measures. The final **Chapter 7** contains the general discussion and direction for future research.

References

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record*, *60*(3), 527.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, *42*(2), 192-212.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological bulletin*, *135*(3), 347-368.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of personality and social psychology*, *69*(6), 1013.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology*, *17*(1), 74-147.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin*, *37*(4), 570-583.

Wittenbrink, B., & Schwarz, N. (Eds.). (2007). *Implicit measures of attitudes*. New York, NY: Guilford Press.

Nederlandstalige Samenvatting

Een psychometrische analyse van keuze-reactietijdmaten

Maten van psychologische attributen, zoals persoonlijkheid en attitudes, spelen een centrale rol in de ontwikkeling van psychologische theorieën en hun dagdagelijkse toepassing. Omdat veel psychologische attributen niet onmiddellijk observeerbaar zijn, zijn metingen veelal gebaseerd op de geobserveerde antwoorden ten aanzien van construct-gerelateerde vragen en dat in een gestructureerde omgeving. Psychologische attributen worden derhalve vaak gemeten aan de hand van vragenlijsten met zelf-rapportage. Het nadeel van zulke vragenlijsten is echter dat deze heel sterk afhankelijk zijn van de bereidheid en bekwaamheid van de respondenten om attributen of gedrag te rapporteren. Sociaal wenselijk of strategisch antwoorden, alsook het niet-bewust zijn van de eigen attitude zijn voorbeelden van welgekende contextfactoren die maten gebaseerd op zelf-rapportage kunnen verstoren (Wittenbrink & Schwarz, 2007).

Als antwoord op deze bezorgdheden werden impliciete maten ontwikkeld, waarvan een groot deel gebaseerd is op keuze-reactietijdtaken. Welgekende voorbeelden van keuze-reactietijdtaken zijn de ‘evaluative priming task’ (Fazio, Jackson, Dunton & Williams, 1995), de ‘Implicit Association Test’ (Greenwald, McGhee & Schwartz, 1998) en de ‘Implicit Relational Assessment Procedure’ (Barnes-Holmes, Barnes-Holmes, Stewart & Boles, 2010). In tegenstelling tot expliciete zelf-rapportage, waarbij participanten directe controle hebben over hun antwoordgedrag, wordt aangenomen dat impliciete maten eerder oncontroleerbare en niet-intentionele antwoordtendensen capteren. Het zijn deze antwoordtendensen die als een proxy voor een specifieke attribuut beschouwd worden.

De introductie van keuze-reactietijdmaten bracht nieuwe psychometrische uitdagingen met zich mee. Deze uitdagingen zijn gerelateerd aan onder meer de betrouwbaarheid en validiteit van deze testen (De Houwer, Teige-Mocigemba, Spruyt & Moors, 2009; LeBel & Paunonen, 2011), de geschiktheid van de gebruikte score-procedures om data, gegenereerd door de testen, te analyseren (Blanton, Jaccard, Gonzales & Christie, 2006) en de identificatie

van factoren die de testresultaten vertekenen (Fiedler, Messner & Bluemke, 2006). Door de alsmaar groeiende toepassingen van keuze-reactietijdmaten in (klinische) assessment situaties, groeit bovendien de nood aan psychometrische modellen van indirecte reactietijdmaten die individuele verschillen in beschouwing nemen. Het is in deze context dat dit proefschrift de psychometrische kwaliteiten van impliciete maten verder heeft onderzocht, verbeterd waar nodig en nieuwe benaderingen heeft ontwikkeld. Om dit te realiseren, werden inzichten van psychometrische modeleringstheorie en nieuwe statistische methodes geïntegreerd.

Hoofdstuk 1 begint met een algemene inleiding over psychometrie en het meten van attributen in psychologie. Vervolgens wordt gereflecteerd over de mogelijke uitdagingen van impliciete maten en worden de algemene doelstellingen van het proefschrift toegelicht.

In **Hoofdstuk 2** worden de 'Probabilistic Index Models' (PIMs) geïntroduceerd als veelbelovende regressiemodellen voor de effectmaat PI of $P(Y_1 < Y_2)$. Waar bij ordinaire lineaire regressie de associatie tussen een uitkomst en een covariaat uitgedrukt wordt als een verschil in gemiddelde, wordt deze associatie nu gedefinieerd in termen van een probabilistische index (PI). Om het begrip PI beter te kunnen bevatten, wordt dieper ingegaan op de betekenis ervan en worden de eigenschappen van de PI als een effectmaat grondig bestudeerd. Zo wordt onder meer aangetoond dat de Wilcoxon–Mann–Whitney test, waarvan de PI beschouwd kan worden als de geassocieerde effectmaat, voor tal van distributies superieur is ten opzichte van de t-test voor wat betreft de power. Verder wordt ook de eigenschap dat de PI invariant is ten opzichte van monotone transformaties uiteengezet. Het is precies deze eigenschap die van de PI een zeer aantrekkelijke maat kan maken binnen de gedragswetenschappen. Wanneer PIMs gebruikt worden, volstaat het immers een monotone relatie te veronderstellen tussen wat geobserveerd wordt en de latente variabele om resultaten op basis van de geobserveerde scores te generaliseren naar het latent construct.

In **Hoofdstuk 3** worden twee alternatieve score-algoritmes voor de 'Implicit Association Test' voorgesteld, waarbij het effect wordt geschat aan de hand van een PIM. Naast het voordeel dat een PI score eenvoudig te interpreteren is, blijkt dat de PI score-algoritmes relatief robuuster zijn tegen de aanwezigheid van extreme observaties ('outliers') en andere eigenschappen van distributies zoals scheefheid, dan klassieke algoritmes zoals de D score-algoritmes. Deze eigenschappen zouden de meetprecisie van de IAT moeten vergroten, wat gereflecteerd wordt in hogere schattingen van betrouwbaarheid. Daarnaast blijkt dat de nieuwe

algoritmes het even goed doen als de bestaande algoritmes voor wat betreft de correlaties met expliciete maten. Met betrekking tot de predictieve validiteit (voor een Politieke IAT) doen de nieuwe algoritmes het iets beter dan de bestaande. Naast deze aantrekkelijke eigenschappen, blijkt het PI score-algoritme een goed alternatief te zijn voor IATs waarbij enkel de tijd van de respons geregistreerd wordt, zonder rekening te houden met de correctheid van de respons (i.t.t. IATs waarbij de tijd van een correcte response gemeten wordt). Er wordt aangetoond dat de fout-correctie, zoals voorgesteld wordt door het D4 score-algoritme (soms ook aangeduid met D600), de berekende score kan verhogen of verlagen afhankelijk van de ware score. Doordat response tijden als ordinaal worden beschouwd door de PI algoritmes, kunnen de responstijden van fouten gemakkelijk vervangen worden door 10,000ms, de maximaal mogelijke tijd voor een IAT. Het effect van deze correctie zal altijd hetzelfde zijn, ongeacht de ware score: fouten worden beschouwd als trials met de traagste tijd.

In **Hoofdstuk 4** wordt het PI-score-algoritme als alternatief algoritme voorgesteld voor de 'Implicit Relational Assessment Procedure'. In dit hoofdstuk wordt aangetoond hoe de PI-scores en D-scores op individueel niveau sterk kunnen verschillen, terwijl er toch een hoge correlatie mogelijk is tussen beide scores. Daarnaast komen ook de aantrekkelijke eigenschappen van de PI score-algoritmes aan bod: robuustheid, eenvoudige interpretatie, het eenvoudig bekomen van betrouwbaarheidsintervallen en p-waarden, en de mogelijkheid om covariaten toe te voegen. Tenslotte wordt er op gewezen dat al deze eigenschappen de PI effect maat aantrekkelijk maken voor 'single-subject' analyses.

In **Hoofdstuk 5** wordt een algemeen kader voorgesteld om de betrouwbaarheid van impliciete testen te benaderen. In het eerste deel wordt gewezen op het belang van conceptuele duidelijkheid. Vervolgens wordt de nadruk gelegd op het belang van de interactie tussen steekproef, test en context wanneer er over betrouwbaarheid van een test gesproken wordt. Gebaseerd op 'Confirmatory Factor Analysis', wordt een perspectief geboden op betrouwbaarheid, waarmee een duidelijk kader gecreëerd wordt om de drie proxies van betrouwbaarheid, consistentie, equivalentie en stabiliteit te schatten. Binnen dit kader wordt het begrip 'parcels' in meer detail besproken en wordt aangetoond dat de verschillende manieren van het construeren van parcels een impact hebben op de schattingen van betrouwbaarheid. Het hoofdstuk eindigt met richtlijnen om het aantal vrijheidsgraden bij het schatten van betrouwbaarheid te beperken.

In **Hoofdstuk 6** worden de aanbevelingen van LeBel en Paunonen (2011) over betrouwbaarheid van impliciete maten geëvalueerd. De centrale stelling in het artikel van LeBel en Paunonen is dat betrouwbaarheid, statistische power en repliceerbaarheid nauw verbonden begrippen zijn. Door een dominante, eenzijdige focus op het vergroten van de foutenvariantie, lijken deze concepten eenduidig en positief te correleren. In dit hoofdstuk wordt echter aangetoond dat, door de foutenvariantie te fixeren en de variantie van de ware score te wijzigen, een tegenovergestelde relatie wordt geobserveerd. Aan de hand van deze betrouwbaarheidsparadox worden een aantal inconsistenties vastgesteld en aangepast waar nodig.

Het laatste deel, **Hoofdstuk 7**, bevat de algemene discussie met een kritische reflectie over het geleverde werk en voorstellen voor mogelijk toekomstig onderzoek.

Referenties

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record*, *60*(3), 527.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, *42*(2), 192-212.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological bulletin*, *135*(3), 347-368.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of personality and social psychology*, *69*(6), 1013.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology*, *17*(1), 74-147.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin*, *37*(4), 570-583.

Wittenbrink, B., & Schwarz, N. (Eds.). (2007). *Implicit measures of attitudes*. New York, NY: Guilford Press.

Data Storage Fact Sheets

Data Storage Fact Sheet for Chapter 2

% Data Storage Fact Sheet

% Name/identifier study
% Author: Maarten De Schryver
% Date: 20 December 2017

1. Contact details
==========

1a. Main researcher

- name: Maarten De Schryver
- address: Henri Dunantlaan 2, B-9000 Gent
- e-mail: Maarten.DeSchryver@Ugent.be

1b. Responsible Staff Member (ZAP)

- name: Jan De Houwer
- address: Henri Dunantlaan 2, B-9000 Gent
- e-mail: Jan.DeHouwer@Ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
==========
* Reference of the publication in which the datasets are reported:
Chapter 2 of PhD dissertation: Introduction to probabilistic index models: regression models for the effect size P (Y1 < Y2)

* Which datasets in that publication does this sheet apply to?:
BtheB, Illustration section

3. Information about the files that have been stored
==========

3a. Raw data
---

* Have the raw data been stored by the main researcher? [ ] YES / [x] NO
If NO, please justify:
The original study is reported in Proudfoot et al. (2003) and part of the data are available in the R package of Hothorn and Everitt.

* On which platform are the raw data stored?
 - [ ] researcher PC
 - [ ] research group file server
 - [X] other (specify): R package HSAUR3

* Who has direct access to the raw data (i.e., without intervention of another person)?
 - [ ] main researcher
 - [ ] responsible ZAP
 - [ ] all members of the research group
 - [ ] all members of UGent
 - [X] other (specify): everyone using R

3b. Other files
---

* Which other files have been stored?
 - [ ] file(s) describing the transition from raw data to reported results. Specify: ...
 - [ ] file(s) containing processed data. Specify: ...
 - [ ] file(s) containing analyses. Specify: ...
 - [ ] files(s) containing information about informed consent
 - [ ] a file specifying legal and ethical provisions
 - [ ] file(s) that describe the content of the stored files and how this content should be interpreted. Specify: ...
 - [X] other files. Specify An R markdown document with all R code can be found in Appendix 1

* On which platform are these other files stored?
 - [X] individual PC
 - [X] research group file server
 - [ ] other: ...

* Who has direct access to these other files (i.e., without intervention of another person)?
 - [X] main researcher
 - [X] responsible ZAP
 - [X] all members of the research group
 - [ ] all members of UGent
 - [ ] other (specify): ...

4. Reproduction
==========
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
  - name:
  - address:
  - affiliation:
  - e-mail:

Data Storage Fact Sheet for Chapter 3

% Data Storage Fact Sheet

% Name/identifier study
% Author: Maarten De Schryver
% Date: 20 December 2017

1. Contact details
==========

1a. Main researcher

- name: Maarten De Schryver
- address: Henri Dunantlaan 2, B-9000 Gent
- e-mail: Maarten.DeSchryver@Ugent.be

1b. Responsible Staff Member (ZAP)

- name: Jan De Houwer
- address: Henri Dunantlaan 2, B-9000 Gent
- e-mail: Jan.DeHouwer@Ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
==========
* Reference of the publication in which the datasets are reported:
Chapter 3 of PhD dissertation: The Probabilistic Index: A new effect size measure for the IAT

* Which datasets in that publication does this sheet apply to?:
“Attitudes 3.0” data set

3. Information about the files that have been stored
==========

3a. Raw data
---

* Have the raw data been stored by the main researcher? [] YES / [X] NO
If NO, please justify: we re-analyzed data from a study reported by BarAnan and Nosek (2014)

* On which platform are the raw data stored?
  - [ ] researcher PC
  - [ ] research group file server
  - [X] other (specify): Open Science Framework: https://osf.io/qf9jx/

* Who has direct access to the raw data (i.e., without intervention of another person)?
  - [ ] main researcher
  - [ ] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [X] other (specify): everyone

3b. Other files
---

* Which other files have been stored?
  - [ ] file(s) describing the transition from raw data to reported results. Specify: ...
  - [] file(s) containing processed data. Specify: ...
  - [X] file(s) containing analyses. Specify: R-script
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how this content should be interpreted. Specify: ...
  - [] other files.

* On which platform are these other files stored?
  - [X] individual PC
  - [X] research group file server
  - [ ] other: ...

* Who has direct access to these other files (i.e., without intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

4. Reproduction
==========
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
  - name:
  - address:
  - affiliation:
  - e-mail:

Data Storage Fact Sheet for Chapter 4

% Data Storage Fact Sheet

% Name/identifier study
% Author: Maarten De Schryver
% Date: 20 December 2017

1. Contact details
==========

1a. Main researcher

- name: Maarten De Schryver
- address: Henri Dunantlaan 2, B-9000 Gent
- e-mail: Maarten.DeSchryver@Ugent.be

1b. Responsible Staff Member (ZAP)

- name: Jan De Houwer
- address: Henri Dunantlaan 2, B-9000 Gent
- e-mail: Jan.DeHouwer@Ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
==========
* Reference of the publication in which the datasets are reported:
Chapter 4 of PhD dissertation: The PI_{IRAP}: An alternative scoring algorithm for the IRAP

* Which datasets in that publication does this sheet apply to?:
gender stereotypes IRAP dataset raw.csv

3. Information about the files that have been stored
==========

3a. Raw data
---

* Have the raw data been stored by the main researcher? [ ] YES / [x] NO
If NO, please justify:
To illustrate, we use an IRAP that was designed to assess gender stereotypes. (see Cartwright, Hussey, Roche, Dunne, & Murphy, 2017)
* On which platform are the raw data stored?
  - [ ] researcher PC

- [ ] research group file server
- [X] other (specify): Cartwrights PC

* Who has direct access to the raw data (i.e., without intervention of another person)?
- [ ] main researcher
- [ ] responsible ZAP
- [ ] all members of the research group
- [ ] all members of UGent
- [X] other (specify): Cartwright

3b. Other files
---

* Which other files have been stored?
- [ ] file(s) describing the transition from raw data to reported results. Specify: ...
- [] file(s) containing processed data. Specify: ...
- [X] file(s) containing analyses. Specify: R script
- [ ] files(s) containing information about informed consent
- [ ] a file specifying legal and ethical provisions
- [ ] file(s) that describe the content of the stored files and how this content should be interpreted. Specify: ...
- [ ] other files. Specify: ...

* On which platform are these other files stored?
- [X] individual PC
- [X] research group file server
- [X] other: Open Science Framework: https://osf.io/4cmsm/

* Who has direct access to these other files (i.e., without intervention of another person)?
- [X] main researcher
- [X] responsible ZAP
- [X] all members of the research group
- [ ] all members of UGent
- [ ] other (specify): Open Science Framework: https://osf.io/4cmsm/

4. Reproduction
==========
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
- name:
- address:
- affiliation:
- e-mail:

Data Storage Fact Sheet for Chapter 5

% Data Storage Fact Sheet

% Name/identifier study
% Author: Maarten De Schryver
% Date: 20 December 2017

1. Contact details
==========

1a. Main researcher

- name: Maarten De Schryver
- address: Henri Dunantlaan 2, B-9000 Gent
- e-mail: Maarten.DeSchryver@Ugent.be

1b. Responsible Staff Member (ZAP)

- name: Jan De Houwer
- address: Henri Dunantlaan 2, B-9000 Gent
- e-mail: Jan.DeHouwer@Ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
==========
* Reference of the publication in which the datasets are reported:
Chapter 5 of PhD dissertation: On the Reliability of Implicit Measures:
Current Practices and Novel Perspectives

* Which datasets in that publication does this sheet apply to?:
“Attitudes 3.0” data set

3. Information about the files that have been stored
==========

3a. Raw data
---

* Have the raw data been stored by the main researcher? [] YES / [X] NO
If NO, please justify: we re-analyzed data from a study reported by BarAnan and Nosek (2014)

* On which platform are the raw data stored?
  - [ ] researcher PC
  - [ ] research group file server
  - [X] other (specify): Open Science Framework: https://osf.io/qf9jx/

* Who has direct access to the raw data (i.e., without intervention of another person)?
  - [ ] main researcher
  - [ ] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [X] other (specify): everyone

3b. Other files
---

* Which other files have been stored?
  - [ ] file(s) describing the transition from raw data to reported results. Specify: ...
  - [] file(s) containing processed data. Specify: ...
  - [X] file(s) containing analyses. Specify: R-script
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how this content should be interpreted. Specify: ...
  - [ ] other files.

* On which platform are these other files stored?
  - [X] individual PC
  - [X] research group file server
  - [ ] other: ...

* Who has direct access to these other files (i.e., without intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

4. Reproduction
==========
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
  - name:
  - address:
  - affiliation:
  - e-mail:

Data Storage Fact Sheet for Chapter 6

% Data Storage Fact Sheet

% Name/identifier study
% Author: Maarten De Schryver
% Date: 20 December 2017

1. Contact details
==========

1a. Main researcher

- name: Maarten De Schryver
- address: Henri Dunantlaan 2, B-9000 Gent
- e-mail: Maarten.DeSchryver@Ugent.be

1b. Responsible Staff Member (ZAP)

- name: Jan De Houwer
- address: Henri Dunantlaan 2, B-9000 Gent
- e-mail: Jan.DeHouwer@Ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
==========
* Reference of the publication in which the datasets are reported:
Chapter 6 of PhD dissertation: Unreliable Yet Still Replicable: A Comment on LeBel and Paunonen (2011)

* Which datasets in that publication does this sheet apply to?:
//

3. Information about the files that have been stored
==========

3a. Raw data
---

* Have the raw data been stored by the main researcher? [ ] YES / [x] NO
If NO, please justify:
We only simulated data, just for illustrative purposes.

* On which platform are the raw data stored?
  - [ ] researcher PC
  - [ ] research group file server
  - [ ] other (specify):

* Who has direct access to the raw data (i.e., without intervention of another person)?
  - [ ] main researcher
  - [ ] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): everyone

3b. Other files
---

* Which other files have been stored?
  - [ ] file(s) describing the transition from raw data to reported results. Specify: ...
  - [ ] file(s) containing processed data. Specify: ...
  - [X] file(s) containing analyses. Specify: R script
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how this content should be interpreted. Specify: ...
  - [ ] other files. Specify:...

* On which platform are these other files stored?
  - [X] individual PC
  - [X] research group file server
  - [ ] other: ...

* Who has direct access to these other files (i.e., without intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

4. Reproduction
==========
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
  - name:
  - address:
  - affiliation:
  - e-mail: