Journal of Experimental Social Psychology 48 (2012) 879–891

Contents lists available at SciVerse ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Reliability of the Go/No Go Association Task

Ben J. Williams [a,*], Leah M. Kaufmann [b]

[a] Faculty of Life and Social Sciences, Swinburne University of Technology, Hawthorn 3122, Australia
[b] School of Psychology, Australian Catholic University, Fitzroy 3065, Australia

ARTICLE INFO

Article history:
Received 20 August 2010
Revised 28 February 2012
Available online 8 March 2012

Keywords:
Implicit measure
Reliability
GNAT
Social cognition
Resampling

ABSTRACT

The Go/No Go Association Task (GNAT; Nosek & Banaji, 2001) is an implicit measure with broad application in social psychology. It has several conceptual strengths to recommend it over other implicit methods, but the belief that it has poor reliability coupled with the absence of a method for calculating this important psychometric property has hindered its wider acceptance and use. Using data obtained from six GNAT studies covering a wide range of content areas, Study 1 compares the properties of different methods for estimating reliability of the GNAT. Study 2 demonstrates a resampling procedure to investigate how reliability varies as a function of block length. Study 1 shows that with appropriately chosen stimuli the GNAT can be a very reliable measure, while Study 2 indicates that as an empirical rule of thumb 50 to 80 trials per block should yield adequate to very good reliability. However, researchers are urged to calculate their own reliability coefficients, to this end we discuss GNAT design issues and provide procedures for calculating GNAT reliability which we hope will enhance the utility of the GNAT as a measure and promote its use in studying implicit cognition.

Introduction

The Go/No Go Association Task (GNAT; Nosek & Banaji, 2001) is a relatively new and promising measure of implicit associations. Although the GNAT has several advantages over other "implicit measures" (e.g., De Houwer & Moors, 2007) it has not been widely used. This seems to be due largely to the lack of a procedure for calculating the reliability of a GNAT, and concerns that the GNAT's reliability is too low for it to be useful. This paper seeks to redress these issues. This paper reviews current practices and outlines conceptual problems in calculating the reliability of the GNAT, empirically examines the properties of some currently (under-) used methods, proposes a resampling method for estimating GNAT reliability, demonstrates that GNATs can achieve good reliability, and gives some guidelines for GNAT design from a reliability perspective.

The value of implicit measures

Like unobtrusive (e.g., Fazio, Jackson, Dunton, & Williams, 1995), nonreactive (e.g., McConahay, 1986), and indirect measures (e.g., Hammond, 1948), implicit measures are designed to be less susceptible than explicit measures to confounds such as impression management (Schlenker, Bonoma, & Tedeschi, 1973), self-presentation concerns (Doherty & Schlenker, 1991), and social desirability (Crowne & Marlowe, 1960; see Fazio & Olson, 2003 for discussion of these issues). Implicit measures have the additional advantage of not requiring introspective awareness, and can therefore be used to assess cognitions (Nisbett & Wilson, 1977) or perceptions (Sanderson & Cantor, 2001) which are not consciously accessible.

The GNAT assesses implicit associations between a target and an attribute. From a participant's viewpoint, a GNAT-measured racial attitude involves completing at least two blocks of trials: one in which trials containing an African American face or a positive word (e.g., "kind") are the targets and should be responded to (e.g., press the spacebar), and a second block where trials containing an African American face or a negative word (e.g., "nasty") are the targets which should be responded to. Both blocks also contain distracter trials (e.g., Caucasian faces and attribute words of the opposite valence), to which participants should make no response, allowing the trial to timeout. Trials terminate when a response is made or the "timeout" period elapses, typically in less than 1 s, to limit the influence of conscious processing. The strength of the association between target and attribute is quantified by the signal detection theory (SDT) measure d'. Using the "equal variance" formula (e.g., Wickens, 2002) d' is calculated as the probit of the proportion of correct responses on target present trials ("hits" in SDT parlance) minus the probit of the proportion of incorrect responses on target absent trials ("false alarms").

Implicit measures can reduce response biases while maintaining the face validity of the study (e.g., the construct of interest need not be obscured), owing to the non-trivial relationship between the construct (e.g., implicit attitude) and method used to asses it (e.g., response speed or accuracy). The transparency and complexity of

[*] Corresponding author.
E-mail address: bwilliams@swin.edu.au (B.J. Williams).

0022-1031/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jesp.2012.03.001

implicit measures give rise to three further advantages. First, the transparency ensures that the information being processed is relevant (for a discussion see De Houwer & De Bruycker, 2007). Second, participants can experience the relative ease or difficulty of responding to certain pairings which can lead to awareness of a previously unknown attitude[1] (e.g., Greenwald, Nosek, & Sriram, 2006). Finally, the task complexity limits the effects of intentional responding (e.g., it is unclear what or how to fake without experience or instructions; e.g., Cvencek, Greenwald, Brown, Snowden, & Gray, 2010; Fiedler & Bluemke, 2005; Kaufmann & Haslam, submitted for publication; Steffens, 2004).

The GNAT

The Go/No Go Association Task (GNAT; Nosek & Banaji, 2001) is conceptually similar to the dominant implicit measure, the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) which currently has more than 1700 citations (Scopus [database], 2012). However, the GNAT has several major methodological advantages over the IAT. The GNAT can assess the implicit association between a single target and attribute (e.g., African American and positive separate from African American and negative) rather than the relative measure produced by the IAT (e.g., African American and positive and Caucasian and negative, compared to American and negative and Caucasian and positive). From a design perspective, this means that the GNAT requires the creation of only one set of stimuli representing the target category, whereas the IAT requires two sets of equivalent (e.g., category size, salience) contrasting stimuli, only one of which actually represents the target construct. Finally, because the GNAT is scored using signal detection parameters (d' and β; see Green & Swets, 1966), this measure allows researchers to assess a participant's implicit associations and response criterion independently (Nosek & Banaji, 2001; see Brendl, Markman, & Messner, 2001 for discussion of the importance of separating these).

Despite the often cited advantages of the GNAT over the IAT, this measure has received little attention and even less use. Of those articles that acknowledge the methodological advantages of the GNAT less than 15% also use the measure, however, those that do have found it a powerful tool. For example, the GNAT has been used to study implicit attitudes to simple attitude objects (e.g., bugs and fruit; Nosek & Banaji, 2001), more complex social groups including gender (Mitchell, Nosek, & Banaji, 2003), racial groups (e.g., Kaufmann & Johnson, 2011; Mitchell et al., 2003; Nosek & Banaji, 2001), and stigmatized groups (e.g., Bassett & Dabbs, 2005; Ranganath, Smith, & Nosek, 2008), as well as implicit stereotyping (e.g., Allen, Sherman, Conrey, & Stroessner, 2009; Blair, Ma, & Lenton, 2001; Gonsalkorale, von Hippel, Sherman, & Klauer, 2009), and implicit prejudice (e.g., Sherman, Stroessner, Conrey, & Azam, 2005; Smith, Stewart, Myers, & Latu, 2008). The GNAT has also proved useful to the study of complex and potentially arbitrary attitude objects such as explicitly stigmatized consumables (e.g., genetically modified food; Spence & Townsend, 2006, 2007) and novel stimuli (e.g., nonwords; Kaufmann & Haslam, 2006). Finally, a real benefit of the GNAT is that it is capable of measuring implicit associations which means it is uniquely suited to investigating implicit targets for which there is no equivalent category (e.g., the self). Thus, the GNAT has been embraced by researchers exploring self-related cognitions including self-humanizing (e.g., Loughnan & Haslam, 2007), the self-concept (e.g., Devos, Viera, Diaz, & Dunn, 2007), personality (e.g., Boldero, Rawlings, & Haslam, 2007) and implicit self-esteem (e.g., Boucher, Peng, Shi, & Wang, 2009; Rudolph, Shröder-Abé, Schütz, Gregg, & Sedikides, 2008).

One reason the GNAT may have received little of its due attention is because it is only one of a raft of new methods competing for attention. We, however, believe that it is more likely that the GNAT is overlooked in favor of reaction time-based measures (e.g., IAT, Extrinsic Affective Simon Task; De Houwer, 2003) because there is no established measure of reliability for d'. As a result, researchers using the GNAT are unable to answer reliability-based criticisms, including the argument that putative low reliability is the reason for the lack of correspondence between the GNAT and other implicit measures (e.g., Cunningham, Preacher, & Banaji, 2001).

Reliability and the GNAT

Reliability is an essential property affecting the use and acceptance of measures. The high reliability of the IAT (e.g., Bosson, Swann, & Pennebaker, 2000; Brunel, Tietje, & Greenwald, 2004; Schmukle, Back, & Egloff, 2008) is often the stated reason it is preferred over other implicit measures (e.g., Conner & Barrett, 2005; Greenwald & Nosek, 2001; Rudman & Heppen, 2003; Rudolph et al., 2008; Schnabel, Asendorpf, & Greenwald, 2008). IATs have shown both satisfactory internal consistency and high test–retest stability (Nosek, Greenwald, & Banaji, 2007), leading Nosek and colleagues to conclude that “part of the IAT's acceptance as an implicit measure may be attributable to its achieving greater reliability than other latency-based implicit measures” (p.274). Thus, an essential first step in increasing the GNAT's acceptance is identifying a method for calculating its reliability and demonstrating that GNAT measures have good reliability.

A review of GNAT studies reveals that most researchers have been satisfied in using the GNAT without considering reliability. We speculate that researchers have avoided this essential issue because of the difficulties associated with calculating reliability for test scores not based on a sum of item scores (i.e., d' is a nonlinear function of responses; see also Nosek & Banaji, 2001 for a discussion). However, there are some exceptions.

One approach has been to forego the benefits of the d' scoring method, and instead use reaction times (e.g., Spence & Townsend, 2007) or reaction time-based D scores (e.g., Teachman, 2007) for which simple split-half reliabilities can be calculated. This approach is supported by Nosek and Banaji's (2001) recommendation that “it is likely that use of response latency [as a scoring method] will result in greater internal reliability” (p. 649). However, this method is inappropriate because the GNAT uses trial deadlines, which artificially truncate the range of response times.

A more popular approach to the problem of quantifying GNAT reliability has been to calculate split-half reliabilities for single blocks or aggregates (e.g., “good” minus “bad”). Reported aggregate reliabilities range from $r = .20$ for a sample of 50 participants who completed six blocks of 60 trials (Nosek & Banaji, 2001) to $r = .65$ for 195 participants who completed four blocks each comprising 48 trials (Rudolph et al., 2008). In contrast, single-block reliabilities are a little higher ranging from $r = .52$ for 195 participants who completed 4 blocks each comprising 48 trials (Rudolph et al., 2008) to $r = .76$ for a single block of 96 trials measuring the implicit association between “self” and “extraversion” (Boldero et al., 2007). It is interesting to observe that this pattern is reversed with test–retest reliabilities. For example, Rudolph et al. (2008) found that their GNAT had modest aggregate test–retest reliability of $r = .51$, but a low test–retest reliability for individual blocks of $r < .38$, results comparable to those for an IAT used in the same study.

With the exception of Nosek and Banaji (2001), researchers who have reported reliabilities either fail to recognize the problems of

[1] Assuming a traditional model where an attitude is conceptualized as a “bipolar dimension that reflects the degree of favorable or unfavorable evaluation of an attitude object” (Blanton, Jaccard, Gonzales, & Christie, 2006, p.193), an attitude may be experienced as the ease of a compatible pairing compared to the difficulty of an incompatible pairing.

non-summed scores as suggested by the application of classical reliability approaches without discussion or justification (e.g., Boldero et al., 2007; Rudolph et al., 2008) or implicitly acknowledge and sidestep the problem by using reaction time-based scores (e.g., *D*). Finally, there is disagreement over whether reliability should be calculated for whole GNATs (i.e., multiple blocks) or for individual blocks without any discussion of which is more appropriate to report and why. Our paper examines the properties of some methods of estimating reliability and develops a robust method of estimating reliability which can also be used to study the effects of block length. In doing so, we address the criticism that the GNAT has generally poor reliability, showing GNAT reliability over a range of topics.

Test theory and considerations for estimating GNAT reliability

In this paper we have adopted an approach to reliability which will be familiar to readers who have studied Classical Test Theory (CTT; e.g., Lord & Novick, 1968). While CTT is not without limitations, we have chosen it because it is the most widely used theory by researchers and practitioners and CTT-based reliability coefficients are routinely supplied with nearly all published tests. Reliability is important in research because, as reliability decreases the observable correlation between two measures is attenuated, regardless of the true correlation between the constructs that the tests measure.

Some have argued that the weak relationship observed between implicit and explicit measures is due to putative low reliability of implicit measures (Bosson et al., 2000; Rudolph et al., 2008). Others interpret this weak relationship as evidence that implicit and explicit measures assess distinct concepts (e.g., Wilson, Lindsey, & Schooler, 2000). Unless the reliability of a GNAT is known, little can be concluded about the meaning of correlations between it and other measures. To date, GNAT reliability has not been systematically studied.

A full explanation of CTT is beyond the scope of this paper (see Allen & Yen, 2002; Traub, 1994 for accessible texts on this theory and its applications), and we assume that the reader is familiar with CTT's partitioning of observed variance into "true" and "error" components. If the true score is stable, the error is uncorrelated with the true score, and errors are uncorrelated across administrations of a measure, reliability can be shown to be equivalent to the correlation between scores on two "instances" of the test, ideally two independent trials of the same test taken by the same cohort (e.g., Guttman, 1945). Practical problems with obtaining two such *independent* "instances" lead to the development of methods which approximate two "equivalent" instances under certain assumptions. Below, we briefly review the appropriateness of these methods for assessing GNAT reliability.

Test–retest reliability correlates two administrations of the same test on a cohort. This method treats only score differences over time as error (which is why it is sometimes called a stability coefficient) and does not account for item sampling effects or person effects such as learning. Implicit tests are prone to substantial practice effects because of the novel nature of the task, rapid learning of the response format and stimuli sets, and priming effects. A repeated GNAT administration is, thus, not properly an independent trial. This coefficient does not capture the reliability of GNATs as they are usually used, but may still be appropriate for investigating the construct validity of "implicit traits."

One way of avoiding practice effect contamination is developing *parallel* or *alternate-forms* of the test. While there are some different GNAT implementations (e.g., the standard computerized GNAT, Nosek & Banaji, 2001; the portable GNAT, Bassett & Dabbs, 2005) the differences are too trivial to be meaningfully considered alternate forms. Furthermore, because implicit tests are rather sensitive to item properties (e.g., word valence, length, and frequency, item pool size, e.g., Bluemke & Friese, 2006) it is generally impractical to construct parallel tests using different items from the same content domain.

The *split-half* method – cutting a test in half to create two tests and correlating these – is currently the most widely used method of calculating GNAT reliability. This avoids some problems associated with the preceding methods. Specifically it can address the practice problem associated with retest methods by avoiding the retest, and depending on how the test is divided (e.g., "odd/even" splits which select alternative trials across the whole block ensures equal distribution of novel and learned trials in each half of the split, whereas a first half–second half division does not) can account for "local" practice. Unfortunately the test halves necessarily correlate less strongly than two tests of the intended length would, and there is a long history of correction formulae (e.g., Spearman, 1910) entailing various assumptions to ameliorate this bias (an issue we return to later).

Cronbach's alpha. In exploring problems associated with calculating reliabilities based on split halves, Cronbach (1951) developed alpha, a lower-bound estimate of reliability that has the interesting property of being algebraically equivalent to the average of all possible split half correlations corrected for test length, often called *internal consistency*. The conceptual appeal of this statistic is that by averaging every possible split, idiosyncratic grouping and sequence effects are smoothed out yielding a more stable picture of how the whole test behaves. Of the methods discussed, this seems the most appropriate for the GNAT.

Algebraic problems associated with calculating GNAT reliability

Why not use an odd/even split and correct for test length? Given the desirable properties of alpha, why not use alpha? Unfortunately, the formulae for both alpha and correction for test length (e.g., the Spearman–Brown "prophecy" formula) cannot be used on a GNAT. These formulae rely on the algebraic properties of item means, variances and covariances (namely the distributive property of expectations) that do not hold for d'. To spell this out, let S be a function that "scores" responses to a set of test items X by summing them (whether the items are equally or differently weighted). Let D be the function for "scoring" GNAT items, that is, the equal variance SDT formula. If we cut test X into two halves comprising sets of items a and b then $S(a+b)=S(a)+S(b)$. Using the distributive law of expectations it is easy to show that the mean of the whole test is equal to the sum of the mean of the halves, and that the variance of the whole test is sum of the variance of each half plus twice the covariance between the two halves. Using these results, many interesting and useful properties of the behavior of test scores can be deduced, including the formula for Cronbach's alpha. None of these properties hold for D.

In general, because d' is a non-linear function of aggregated item scores, $D(a+b)\neq D(a)+D(b)$, the exception being when a and b are *exactly* the same. This is because equal variance SDT formula involves the inverse cumulative density function (probit) of the normal distribution, Φ^{-1}, and in general $\Phi^{-1}(a+b)\neq\Phi^{-1}(a)+\Phi^{-1}(b)$. This can be proven, but the reader can confirm this empirically with just a few numeric examples. To complicate matters further, the distribution of d' is intractable (see Miller, 1996 for a study of this problem) and there is no neat algebraic result regarding d' that we can exploit to even approximate statistics like alpha. This leaves researchers contemplating the appropriateness of different split-half procedures.

In what follows we present two studies. The first explores the effect of different splitting strategies on reliability estimates. This provides a basis for comparing results from published studies. The second study demonstrates a sampling with replacement procedure which provides an empirical estimate of reliability that controls for order effects and accounts for block length. This procedure also allows examination of the effect of block length on reliability, which can be used to help guide GNAT design.

Study 1: estimating reliability for the GNAT — effects of splitting strategy

We have argued that the test–retest and parallel form methods are inappropriate and impractical. Thus, split-half methods are the only viable option currently available. It is presently unknown how much estimates vary due to the *particular* split chosen (e.g., random, odd/even). Hence, a discussion of *how* to split GNAT blocks is pertinent.

The ideal split should balance item and method idiosyncrasies: in the case that items are words, word length, valence and frequency; in the case of picture items, salience and valence; as well as practice (e.g., novel tasks and stimulus changes between blocks may yield substantially better performance for later parts of the test), fatigue, and sequence effects (e.g., the participant assumes that many signal trials in a row is unlikely and thus gives a "noise" response, a kind of gambler's fallacy common for signal detection tasks, see Wickens, 2002). If item presentation is randomized in a block, an odd/even split would satisfy many of these requirements — because each half contains an equal density of items from each part of the test it accounts for some practice and fatigue effects. However, this does not fully account for item-sampling and local sequence effects. It is for reasons like this that split-half methods are not widely used in contemporary test development (i.e., choosing splits is fraught with conceptual issues) and better methods exist for some types of tests. We conclude that their use with GNATs is likely due to the lack of a viable alternative (discussed below).

An internal consistency method similar to Cronbach's alpha might be a better choice because it avoids the problem of relying on a single split to estimate reliability, but no such method exists for the GNAT. The oft-cited appeal of alpha in many testing contexts is that it yields the average of all possible split-half correlations. Since we cannot use this shortcut, why not simply calculate this average by brute force? For an N-item test there are $\frac{N!}{2\left[\left(\frac{N}{2}\right)!\right]^2}$ possible split halves. This is "only" 92,378 splits for a 20 trial test, but it is a 23 digit number for a more desirable 80 trial GNAT block which is computationally expensive even by modern standards. So, we suggest that an empirical approximation will suffice. Instead of computing every possible split-half correlation, a good approximation of this statistic can be obtained from the average of a sufficiently large number (say 1000) of random splits. This method should yield a statistic that has many of the properties of alpha, but is not corrected for test length (which alpha is, if items are tau-equivalent) and, will thus, consistently underestimate the true value. For brevity, we will refer to this process of generating the splits the Random Sample of Split Halves (RaSSH).

Since there is no readily available software package for computing GNAT reliability researchers must obtain split-half reliability estimates manually by recoding their datasets to obtain two scores for each participant and then correlating them. To date, only three papers have reported reliability for GNAT d' scores and these have been based on either a first-half/second-half or an odd/even split. As already noted, the former is not recommended because it does not take sequence effects into account. The latter is defensible and should actually be a good choice since the effects of practice and fatigue should be equally present in the two halves. The odd/even procedure has to be adapted for those GNATs that alternate the presentation of category and attribute trials (e.g., Ranganath et al., 2008),[2] otherwise odd–even splits would end up with only one type of item (e.g., category or attribute) in each half. In these cases, adjacent *pairs* of trials should always be kept together in a split *or* splits be made at random to avoid confounding test half with stimulus type. The sensitivity of odd/even reliability estimates to other sequence effects is not known. For this reason (and because random item sampling deals with both random and alternating trial type GNATs) we argue that the RaSSH mean and distribution add critical information about a GNAT's sensitivity to several sources of potential error and is, therefore, currently the most useful reliability measure for the GNAT.

The RaSSH means should have similar values to odd/even splits for GNATs whose reliability is relatively unaffected by practice effects. However, unlike many psychological measures, GNAT items are usually randomized across participants, therefore odd/even splits will contain different items and item sequences for each participant. In contrast, resampling will average out idiosyncratic sequence effects and, thus, it could be argued that this makes the RaSSH a better indicator of GNAT reliability when generalizing from a research sample where each new case will get items in a different order (although it is not necessarily the best indicator for a single case). Finally, large discrepancies between odd/even split and RaSSH mean or high variability in the RaSSH distribution would be diagnostic of severe sensitivity sequence effects, including practice effects, or problems with item randomization.

Study 1 examined GNAT reliability estimates based on three different split-half correlations: first-half/second-half split, odd/even split, and RaSSH mean. d's in these distributions simulate plausible trial sequences based on empirical distributions from real participants. The RaSSH distribution captures values the reliability could take had the same set of trials (and responses) arisen in a different sequence (e.g., more correct responses towards the end of a block due to practice, random clumping of identical stimuli due to chance). Although we cannot point to a particular split and say that split represents the value of r obtained for some particular level of practice, fatigue or item presentation sequence, the value of r that would result from that level of practice, fatigue, etc. is somewhere in the distribution. Thus the extrema and shape of the RaSSH distribution indicate a GNAT's possible sensitivity to practice, fatigue, sequence, and item sampling effects. By comparing the first-half/second-half and odd/even split to the RaSSH distributions we can comment on the properties of each method and discuss their suitability for computing the reliability of the GNAT. In the process of computing the first-half/second-half split reliabilities we can also estimate the effects of learning or fatigue over the course of a GNAT block by comparing d' for each half. Since this study examines data collected on a diverse range of constructs, it also provides a good indication of the range of values GNAT reliability can be expected to have. Thus, we can determine whether the GNAT's poor reception is deserved, or whether a combined lack of understanding and inappropriate statistical analyses have led to it being unfairly underrated.

Method

Participants

Participants were 276 undergraduate psychology students and 134 members of the general public who each participated in only one of six studies. Participants who took part in the attitude studies were: 44 undergraduate students who completed an implicit attitudes to bugs and fruit GNAT including 16 men and 28 women with a mean age of 18.50 years ($SD = 1.07$); 90 undergraduate students who completed an implicit racism GNAT including 16 men and 74 women with a mean age of 22.98 years ($SD = 10.01$); 49 undergraduate students who completed an implicit ageism GNAT including 8 men and 41 women with a mean age of 21.00 years ($SD = 6.54$); and 134 members of the general public who competed an implicit homophobia GNAT including 67 males, and 59 females with a mean age of 24.83 years ($SD = 7.91$). Fifty-five participants took part in a study of the implicit associations underpinning the folk psychiatry model (Haslam, 2005) including 12 males and 43 women aged from 17 to 35 years of age ($M = 22.32$, $SD = 7.39$). Finally, 67 participants took part in a study examining the role of familiarity on implicit

[2] Note that the data we report was gathered entirely from designs with random item presentation orders.

evaluations including 16 men and 51 women aged from 17 to 41 years ($M = 20.45$, $SD = 5.34$).

Materials and design

Implicit attitude GNATs: bugs–fruit, ageism, racism, and homophobia. All participants completed GNATs as part of larger studies. Implicit attitudes to bugs and fruit were assessed by a four-block GNAT (Kaufmann, 2010) pairing bugs (e.g., ant, butterfly) or fruit (e.g., banana, apple), with positive (e.g., love, happy) or negative (e.g., hate, nasty) attributes as targets in a block. Each target and distracter within the block was represented by twenty stimuli. Implicit racism was assessed by a four-block GNAT (Kaufmann & Johnson, 2011) pairing dark skin-tone faces or light skin-tone faces (a stimuli set developed by Nosek et al., 2007), with positive (e.g., love, happy) or negative (e.g., hate, nasty) attributes as targets in a block. Each target and distracter within the block type was represented by 6 stimuli. Both implicit ageism and implicit homophobia were assessed by two-block GNATs (Anderson & Kaufmann, 2011; Kaufmann, 2010) pairing old faces (a stimuli set developed by Nosek et al., 2007) or faces of famous gay men (e.g., Elton John), with positive (e.g., love, happy) or negative (e.g., hate, nasty) attributes as targets in a block. Category distracters in the ageism GNAT were young faces (Nosek et al., 2007) and photos of similarly aged and equally famous straight men in the homophobia GNAT (e.g., Jon Bon Jovi). Each target and distracter within the block type was represented by 6 stimuli.

Blocks in the bugs–fruit and racism studies consisted of 20 practice trials and 100 experimental trials, ageism blocks consisted of 8 practice trials and 40 experimental trials, and blocks in the homophobia consisted of 10 practice trials and 80 experimental trials. Blocks in all studies comprised equal numbers of target and distracter trials. Practice stimuli were a subset of experimental stimuli. In four-block GNATs, all stimuli were used as targets and distracters (e.g., the fruit and good block included fruit and good stimuli as targets, and bugs and bad stimuli as distracters). Trials were terminated when a response was made or after a 700 ms timeout. Participants received feedback on each trial (a red “X” for incorrect responses or a green “O” for correct responses presented for 350 ms following the trial), and an intertrial interval of 100 ms followed feedback. Each block was separated by a short break. Participants were informed what the targets for a block would be immediately before commencing that block.

Folk psychiatry. As part of a larger study (Kaufmann & Haslam, 2007) a two-block GNAT assessed the implicit associations between the target category “mental disorder”, represented by 10 recognizable mental disorders (e.g., depression, schizophrenia) and the attributes normal and abnormal, represented by 10 terms each (e.g., regular, everyday, versus unusual, deviant) theorized to form the two poles of a of pathologizing dimension (Haslam, 2005). A second nine-block GNAT assessed the implicit associations between clusters of mental disorders (i.e., medicalized, moralized, and psychologized mental disorders; see Haslam, Ban, & Kaufmann, 2007 for a discussion) and the three cognitive factors hypothesized to characterize lay thinking about mental disorders (i.e., medicalizing, moralizing, and psychologizing; Haslam, 2005). Again, 10 stimuli represented each category and attribute.

The two-block GNAT blocks comprised 8 trials and 68 experimental trials. The nine-block GNAT blocks comprised 20 practice trials and 100 experimental trials. All blocks had equal numbers of target and distracter trials. All stimuli were used as targets and distracters. Trials were terminated when a response was made or after a 1000 ms timeout. All other aspects of procedure were the same as the bugs and fruit study.

Familiarity. As part of a larger study (Kaufmann & Haslam, 2006), a four-block GNAT assessed the implicit associations between familiar nonwords (i.e., ending in “ald”, for familiarization procedure see Kaufmann, 2010) or unfamiliar nonwords (i.e., ending in “ert”), represented by six pronounceable strings of four to seven letters, and frequent or infrequent positive and negative attributes. Six stimuli that had both highest rated positivity or negativity (Siegle, 1994) and the highest or lowest frequency (Francis & Kucera, 1982) represented each of the attributes. Practice blocks comprised 12 trials and experimental blocks comprised 48 trials including equal numbers of target and distracter trials (i.e., each of the six stimuli was presented twice during the experimental blocks). All stimuli were used as targets and distracters. Trials were terminated when a response was made or after a 700 ms timeout. All other aspects of procedure were the same as the attitudes to bugs and fruit study.

Procedure

All participation was undertaken in small supervised sessions in a university computer laboratory. Participants received a plain language statement and gave informed consent prior to participation. All tasks were presented on PCs using Inquisit 2.0.60616 [Computer software] (2006).

Analyses

All analyses were performed using scripts written by the authors[3] in the R programming language (version 2.7; R Development Core Team, 2008). For each block in each dataset reliability coefficients were obtained by splitting each participant's trials into two equal-sized pools and calculating the Pearson correlation between d' scores for each pool across participants. The d' for each half-test was calculated using the equal variance signal detection theory equations (e.g., Wickens, 2002) with target trials as signal and distracter trials as noise (see Nosek & Banaji, 2001). Only experimental trials were analyzed.

Three different splitting methods were examined. First, odd/even trials splits were created, with half-tests constructed from every second signal (i.e., target) and every second noise (i.e., distracter) trial. Second first-half/second-half splits were created by assigning the first N/2 target trials and N/2 distracter trials to each half-test. Finally, the RaSSH was produced by randomly sampling equal numbers of signal and noise trials without replacement to create two equal sized sub-blocks. This process was repeated 1000 times so that a distribution of the between-half correlations was constructed.

Reliabilities for two-block difference scores were calculated where appropriate pairings could be identified (e.g., combining fruit–good and fruit–bad to yield an overall attitude to fruit). For each participant in each study the distribution of the difference between d' for appropriate pairs of blocks was calculated using the same splitting and subsampling techniques as for single blocks.

Results

Table 1 summarizes the different reliability estimates for the “attitude” GNATs (bugs/fruit, racism, ageism and homophobia). Reliability statistics are summarized in Tables 2 and 3 for the folk psychiatry and familiarity GNATs respectively, while Table 4 shows the results for the two-block difference scores.

Single blocks

The mean d' for the second half of each block was greater than that of the first half in every case, except for the “black–bad” block (see Table 1). These differences were significant at the .05 level in all cases except for the “black–bad”, “white–good” and “old–bad”

[3] The R scripts can be obtained by contacting the authors.

Table 1
Reliability estimates for attitudes to bugs and fruit, racism, ageism, and homophobia GNATs.

Targets	Odd/even split					First half/second half split					RaSSH distribution					
		d′ of half					*d′* of half				Internal consistency (*r*)†					
	Inter-half correlation	Odd half mean (*SD*)	Even half mean (*SD*)	*t*	*p*	Inter-half correlation	1st half mean (*SD*)	2nd half mean (*SD*)	*t*	*p*	Mean (*SD*)	Min	1st quartile	Median	3rd quartile	Max
Bugs–bad	.75	2.04 (0.75)	2.15 (0.67)	1.40	.18	.54	1.96 (0.83)	2.25 (0.62)	2.71	<.01	.65 (0.06)	.44	.61	.66	.70	.85
Bugs–good	.57	1.21 (0.64)	1.26 (0.64)	0.50	.62	.41	0.94 (0.65)	1.60 (0.76)	5.67	<.001	.60 (0.07)	.37	.55	.60	.65	.79
Fruit–bad	.57	1.27 (0.65)	1.40 (0.78)	1.30	.20	.64	1.10 (0.68)	1.57 (0.74)	5.15	<.001	.66 (0.06)	.45	.62	.66	.70	.83
Fruit–good	.57	2.13 (0.63)	2.15 (0.71)	0.19	.85	.44	2.02 (0.62)	2.28 (0.76)	2.42	.02	.58 (0.07)	.34	.54	.59	.63	.81
Black–good	.73	2.68 (0.88)	2.61 (0.89)	1.04	.30	.69	2.32 (0.83)	3.06 (0.94)	9.82	<.001	.73 (0.03)	.63	.71	.74	.76	.85
Black–bad	.83	3.19 (0.84)	3.10 (0.92)	1.70	.09	.67	3.17 (0.97)	3.15 (0.87)	0.29	.78	.80 (0.03)	.71	.79	.80	.82	.87
White–good	.71	2.88 (0.66)	2.86 (0.70)	0.57	.57	.59	2.81 (0.79)	2.92 (0.63)	1.57	.12	.68 (0.05)	.52	.65	.69	.72	.81
White–bad	.60	2.21 (0.09)	2.30 (0.85)	1.03	.31	.56	2.05 (0.92)	2.48 (0.86)	4.90	<.001	.66 (0.04)	.51	.64	.66	.69	.78
Old–good	.48	2.45 (0.57)	2.50 (0.62)	0.39	.70	.40	2.34 (0.51)	2.64 (0.72)	2.10	.047	.47 (0.13)	−.05	.38	.47	.56	.83
Old–bad	.51	2.85 (0.08)	2.89 (0.60)	0.27	.79	.40	2.79 (0.68)	3.01 (0.79)	1.34	.19	.60 (0.01)	.26	.54	.61	.67	.84
Gay–good	.75	1.974 (0.94)	2.12 (0.96)	2.45	.02	.70	1.77 (0.94)	2.30 (1.00)	9.17	<.001	.76 (0.06)	.12	.27	.31	.35	.50
Gay–bad	.69	1.92 (0.92)	1.82 (0.84)	1.66	.10	.65	1.56 (0.83)	2.21 (0.94)	9.96	<.001	.70 (0.03)	.67	.74	.76	.78	.84

Note. Bugs–fruit, $n=44$, 100 trials per block. Racism, $n=90$, 100 trials per block. Ageism, $n=49$, 48 trials per block. Homophobia, $n=134$, 80 trials per block. Absolute t-values given.
RaSSH — Random Sample of Split Halves. † — see text.

blocks. The *d′* for odd/even halves only differed significantly for two of the 33 blocks.

As anticipated, the first-half/second-half splits yielded lower reliability estimates than both the odd/even (30 out of 33 blocks) and RaSSH methods (28 out of 33 blocks; see Tables 1–3). The difference between these estimation methods averaged across all blocks within a study ranged from a relatively small 4.99% for the folk psychiatry data and 6.65% for the homophobia data, to a moderate 14.29% and

Table 2
Reliability estimates for folk psychiatry GNAT as a function of targets.

Targets	Odd/even items					First half/second half split					RaSSH distribution					
		d′ of half					*d′* of half				Internal consistency (*r*)†					
	Inter-half correlation	Odd half mean (*SD*)	Even half mean (*SD*)	*t*	*p*	Inter-half correlation	1st half mean (*SD*)	2nd half mean (*SD*)	*t*	*p*	Mean (*SD*)	Min	1st quartile	Median	3rd quartile	Max
Mental disorders — abnormal^	.78	2.36 (0.94)	2.54 (0.93)	−2.22	.03	.71	2.14 (1.06)	2.76 (0.89)	6.17	<.001	.79 (0.04)	.65	.76	.79	.81	.89
Mental disorders — normal^	.73	1.94 (0.87)	1.86 (1.00)	0.89	.38	.68	1.59 (0.94)	2.23 (0.96)	6.22	<.001	.74 (0.04)	.53	.72	.75	.77	.86
Medicalized mental disorders — medicalizing	.72	2.44 (0.74)	2.39 (0.68)	0.65	.52	.67	2.23 (0.78)	2.64 (0.69)	5.04	<.001	.69 (0.05)	.52	.66	.69	.73	.85
Medicalized mental disorders — moralizing	.80	1.72 (0.96)	1.78 (0.94)	−0.76	.45	.78	1.59 (0.99)	1.94 (0.93)	4.07	<.001	.81 (0.03)	.70	.79	.81	.83	.90
Medicalized mental disorders — psychologizing	.81	2.17 (0.84)	2.19 (0.88)	−0.31	.76	.79	2.11 (0.92)	2.27 (0.85)	2.06	.04	.79 (0.04)	.66	.77	.79	.82	.90
Moralized mental disorders — medicalizing	.73	2.27 (0.84)	2.36 (0.77)	−1.08	.29	.76	2.21 (0.81)	2.41 (0.81)	2.61	.01	.76 (0.04)	.60	.73	.76	.78	.89
Moralized mental disorders — moralizing	.87	1.74 (0.94)	1.82 (1.12)	−1.16	.25	.82	1.68 (1.04)	1.88 (1.03)	2.46	.02	.84 (0.03)	.75	.82	.84	.86	.92
Moralized mental disorders — psychologizing	.76	2.15 (0.81)	2.26 (0.88)	−1.33	.19	.73	2.05 (0.84)	2.36 (0.88)	3.6	<.001	.77 (0.04)	.64	.75	.78	.80	.88
Psychologized mental disorders — medicalizing	.57	2.36 (0.63)	2.36 (0.69)	−0.05	.96	.53	2.25 (0.74)	2.51 (0.61)	2.97	<.005	.60 (0.06)	.36	.56	.61	.65	.82
Psychologized mental disorders — moralizing	.80	1.76 (0.89)	1.83 (0.95)	−0.85	.40	.82	1.58 (0.86)	1.99 (0.96)	5.49	<.001	.79 (0.04)	.68	.77	.79	.82	.88
Psychologized mental disorders — psychologizing	.83	2.26 (0.80)	2.24 (0.83)	0.36	.72	.73	2.08 (0.85)	2.45 (0.86)	4.44	<.001	.78 (0.04)	.64	.75	.78	.80	.87

Note. $n=55$ for all groups, ^number of trials per block = 68, number of trials for all other blocks = 100. Absolute t-values given.
RaSSH — Random Sample of Split Halves. † — see text.

Table 3
Reliability estimates for familiarity GNAT as a function of targets.

Targets	Odd/even items					First half/second half split					RaSSH distribution					
		d′ of half					*d′* of half				Internal consistency (*r*)†					
	Inter-half correlation	Odd half mean (*SD*)	Even half mean (*SD*)	*t*	*p*	Inter-half correlation	1st half mean (*SD*)	2nd half mean (*SD*)	*t*	*p*	Mean (*SD*)	Min	1st quartile	Median	3rd quartile	Max
Practice blocks^ (n = 68)																
Flowers–directions	.58	1.76 (0.79)	1.68 (0.88)	0.58	.56	.51	1.33 (0.87)	2.19 (0.85)	6.17	<.001	.61 (0.07)	.29	.56	.61	.66	.82
Flowers–temperature	.58	2.32 (0.78)	2.21 (0.96)	0.84	.40	.65	1.94 (0.86)	2.63 (0.82)	5.99	<.001	.64 (0.07)	.41	.60	.65	.69	.85
Condition 1 — frequent attributes (n = 33)																
Familiar nonwords — (frequent) positive	.69	2.06 (0.97)	2.00 (0.90)	0.58	.56	.59	1.69 (0.98)	2.41 (0.94)	5.08	<.001	.67 (0.06)	.43	.63	.68	.72	.87
Familiar nonwords – (frequent) negative	.70	1.88 (1.11)	2.05 (1.00)	1.24	.22	.63	1.77 (1.07)	2.17 (1.06)	2.74	<.01	.73 (0.05)	.56	.70	.73	.77	.88
Familiar nonwords — (infrequent) positive	.67	1.94 (0.86)	1.98 (0.92)	0.42	.68	.53	1.63 (1.02)	2.33 (0.83)	4.73	<.001	.64 (0.07)	.35	.59	.65	.69	.84
Familiar nonwords — (infrequent) negative	.62	2.20 (0.96)	2.24 (0.97)	0.29	.77	.60	1.97 (1.02)	2.44 (0.91)	3.38	<.005	.71 (0.06)	.50	.67	.71	.75	.85
Condition 2 — infrequent attributes (n = 34)																
Unfamiliar nonwords — (frequent) positive	.76	2.16 (1.09)	2.10 (0.89)	0.49	.63	.66	1.97 (1.04)	2.27 (1.03)	2.07	<.05	.73 (0.06)	.47	.69	.73	.77	.89
Unfamiliar nonwords — (frequent) negative	.68	1.80 (0.96)	1.75 (0.88)	0.38	.70	.57	1.39 (1.09)	2.19 (0.83)	5.06	<.001	.64 (0.08)	.37	.59	.64	.70	.85
Unfamiliar nonwords — (infrequent) positive	.83	1.80 (1.05)	1.64 (1.04)	1.56	.13	.70	1.44 (1.11)	2.07 (1.07)	4.32	<.001	.73 (0.06)	.51	.70	.74	.77	.89
Unfamiliar nonwords — (infrequent) negative	.64	1.96 (0.88)	1.81 (0.92)	1.19	.24	.46	1.71 (0.98)	2.09 (0.91)	2.27	<.05	.65 (0.07)	.39	.60	.65	.71	.85

Note. ^number of trials per block = 48, all other blocks have 100 trials. Absolute t-values given.
RaSSH — Random Sample of Split Halves. † — see text.

15.29% for the familiarity and racism data respectively, and a substantial 24.13% for the bugs–fruit data. The ageism data showed the largest differences, more than 20% in both blocks.

The RaSSH mean reliabilities were very similar to those for the odd/even splits. The largest difference between RaSSH and odd/even reliabilities for any block was .10 (i.e., “bugs–bad” and “unfamiliar nonwords–infrequent positive”) and the average difference across all studies was only .003 (*SD* = 0.046). In 28 of the 33 blocks odd/even and RaSSH results differed by .03 or less. The mean of RaSSH over all blocks was .70 (*SD* = 0.08), the mean for odd/evens splits was .69 (*SD* = 0.10). The mean of the first half/second half splits over all blocks was substantially lower, .62 (*SD* = 0.10).

As can be seen in Tables 1, 2, and 3, in 21 of the 33 blocks, the mean and median of the RaSSH distribution were identical to the second decimal place. For the remaining twelve blocks, the median was larger than the mean, but only in the second decimal place. Inspection of these results reveals that, in slightly more than half of the cases, the first quartile is further from the median than the third quartile, and in all cases the minimum was slightly further from the median than the maximum, although such differences were mostly in the second decimal place. This indicates that the reliability estimates based on single random splits have a nearly symmetrical distribution with a very slight negative skew.

Two block difference scores

The mean *d′* difference scores did not show a consistent tendency to be higher in the second half of each block pair, most likely due to different learning rates for the blocks in each pair. This is unsurprising given differences in the difficulty between purportedly congruent and incongruent blocks (e.g., Greenwald et al., 2006). Except for this, the pattern of reliability estimates was similar to the single block case: odd/even reliability estimates were generally higher than those based on first-half/second-half split (13 of 17 pairs), and 14 of the 17 pairs had higher mean RaSSH reliabilities than first-half/second-half split reliabilities. In contrast to the pattern for single block data, odd/even reliabilities for difference tended to be higher than RaSSH for difference scores (9 of 17 compared to 7 of 17).

Similar to the single block results, the RaSSH mean and median were identical to the second decimal place in most cases (15 of 17) with the median being higher in the remaining two cases. All RaSSH distributions showed a slight negative skew. Reliability estimates of two block difference scores were lower than their component blocks. The mean reliability across all studies was .38 (*SD* = 0.18), .50 (*SD* = 0.10), and .47 (*SD* = 0.11) for the first/second, odd/even and RaSSH methods respectively.

Discussion

Single blocks

The results indicate that the GNAT can achieve acceptable reliability. While there is no hard and fast rule for how good reliability must be, a widely accepted guideline for Cronbach's alpha is that .60 is “acceptable” (Nunnally, 1967), while values around .80 are considered appropriate for research and a minimum value for clinical applications (Nunnally & Bernstein, 1994). While higher values are generally better, values greater than .90 indicate redundancy in the items (Streiner, 2003). According to RaSSH results, all GNATs had acceptable to very good reliability by these criteria, with the exception of the blocks from the ageism study. This was only true of odd/even split reliabilities if rounding to one decimal place was used, and not true for more than half of the first-half/second-half reliabilities. The results also show that reliability estimates are sensitive to the method of splitting. The results confirm our prediction that the first half/second half split yields substantially lower reliability estimates than the other two methods. Many of these estimates lie in the lower quartiles of the RaSSH distributions which would suggest that they are “worst case” splits. On this basis we conclude that first-half/second-half splits are not appropriate for estimating GNAT reliability.

Table 4
Reliability estimates for two-block difference scores for implicit attitudes, folk psychiatry concepts and familiarity as a function of targets.

Targets and attributes	Odd/even items					First half/second half split					RaSSH distribution					
	d' of half					d' of half					Internal consistency (r)†					
	Inter-half correlation	Odd half mean (SD)	Even half mean (SD)	t	p	Inter-half correlation	1st half mean (SD)	2nd half mean (SD)	t	p	Mean (SD)	Min	1st quartile	Median	3rd quartile	Max
Bugs: bad–good[a]	.33	0.85 (0.64)	0.90 (0.71)	0.44	.66	.04	1.04 (0.80)	0.67 (0.64)	2.49	.02	.24 (0.10)	−.05	.17	.24	.31	.56
Fruit: good–bad[a]	.40	0.89 (0.74)	0.77 (0.81)	0.89	.38	.29	0.95 (0.75)	0.74 (0.92)	1.38	.17	.38 (0.09)	.09	.32	.38	.45	.66
Black: bad–good[b]	.62	0.53 (0.88)	0.51 (1.01)	0.22	.83	.46	0.88 (1.01)	0.11 (0.94)	7.10	<.001	.60 (0.05)	.43	.57	.60	.63	.74
White: good–bad[b]	.39	0.67 (0.92)	0.55 (0.72)	1.25	.22	.26	0.78 (0.91)	0.43 (0.82)	3.09	.003	.40 (0.07)	.17	.36	.40	.45	.60
Old: bad–good[c]	.34	0.41 (0.88)	0.40 (0.68)	0.05	.96	.38	0.45 (0.76)	0.37 (0.85)	0.42	.69	.41 (0.14)	.01	.33	.42	.50	.79
Gay: bad–good[d]	.34	0.58 (1.01)	0.45 (0.97)	0.97	.33	.43	0.25 (0.96)	0.87 (0.93)	6.96	<.001	.43 (0.05)	.27	.39	.43	.46	.59
MD ab–norm[e]	.56	0.44 (0.94)	0.71 (1.17)	1.98	.053	.49	0.56 (1.17)	0.56 (1.02)	0.00	–	.60 (0.06)	.41	.56	.61	.65	.75
Medicalizing med–mor[f]	.57	0.77 (0.89)	0.65 (1.06)	0.95	.348	.66	0.66 (1.03)	0.76 (0.85)	0.95	.35	.58 (0.06)	.37	.54	.58	.63	.79
Medicalizing med–psy[f]	.54	0.28 (0.65)	0.20 (0.83)	0.79	.44	.51	0.11 (0.77)	0.40 (0.74)	2.88	.006	.38 (0.08)	.14	.32	.38	.43	.65
Moralizing: mor–med[f]	.55	−0.55 (0.81)	−0.58 (1.03)	0.31	.76	.52	−0.56 (1.01)	−0.55 (0.84)	0.10	.92	.54 (0.07)	.31	.49	.54	.59	.74
Moralizing: mor–psy[f]	.70	−0.42 (1.01)	−0.48 (0.90)	0.52	.61	.53	−0.38 (1.05)	−0.52 (0.95)	1.05	.30	.60 (0.06)	.39	.56	.60	.65	.79
Psychologizing psy–med[f]	.51	−0.09 (0.85)	−0.12 (0.95)	0.27	.79	.44	−0.20 (0.96)	−0.01 (0.88)	1.47	.15	.51 (0.07)	.25	.47	.51	.56	.74
Psychologizing psy–mor[f]	.46	0.56 (0.87)	0.45 (0.95)	0.85	.40	.49	0.52 (0.86)	0.51 (0.99)	0.07	.94	.53 (0.07)	.29	.48	.53	.58	.71
Fam and freq: pos–neg[g]	.41	0.18 (0.95)	−0.05 (0.98)	1.36	.18	.13	−0.08 (1.05)	0.23 (1.02)	1.40	.17	.38 (0.10)	.09	.31	.38	.45	.66
Fam and inf: pos–neg[g]	.52	0.36 (0.92)	0.35 (0.75)	0.07	.94	.08	0.58 (1.13)	0.08 (0.90)	2.10	.04	.33 (0.11)	.02	.25	.33	.41	.69
Unf and freq: neg–pos[g]	.48	0.26 (0.90)	0.25 (1.16)	0.06	.96	.35	0.34 (1.22)	0.12 (0.88)	1.11	.27	.49 (0.09)	.19	.43	.49	.56	.78
Unf and inf: neg–pos[g]	.50	0.16 (1.03)	0.17 (0.90)	0.05	.96	.29	0.27 (1.07)	0.02 (0.98)	1.19	.24	.41 (0.10)	.03	.34	.41	.47	.70

Note. Absolute t-values are shown. Ab — abnormal; norm — normal; med — medical; mor — moral; psy — psychological; fam — familiar; unf — unfamiliar; freq — frequent; inf — infrequent; neg — negative; pos — positive.
RaSSH — Random Sample of Split Halves. † — see text.
[a] $n=44$ for all groups, 100 trials per block.
[b] $n=90$, 100 trials per block.
[c] $n=49$, 48 trials per block.
[d] $n=134$, 80 trials per block.
[e] $n=55$, 68 trials per block.
[f] $n=55$, 100 trials per block.
[g] $n=67$, 100 trials per block.

Furthermore, the consistently larger scores on the second half of each block are evidence of substantial learning effects. Researchers are urged to consider practice and block length effects in their GNAT designs.

While some very high and very low correlations were observed at the extremes of the RaSSH distributions, the distributions were close to symmetric about the mean (and median) and displayed relatively small spread. This pattern of results indicates that while using a single split to calculate reliability could result in substantial over- or under-estimates, a single random split will generally yield an unbiased estimate of the mean of all split halves.

Perhaps surprisingly, the odd/even splitting method yielded very similar reliability estimates to the RaSSH averages, although overall the RaSSH estimates were marginally higher and varied less within a study. While an odd/even split *could* result in substantial error (as evidenced by the tails of the RaSSH distribution), these results suggest that this has not occurred for the datasets under consideration. It seems that the odd/even split is an acceptable method, but the RaSSH should be preferred, being less variable and more robust to sampling anomalies.

As expected, reliability varied as a function of GNAT content and block length. With the exception of the “fruit–good” and “old–good” blocks, the RaSSH reliability estimates were all above the “acceptable”.60 cutoff and the average reliability across all datasets was good, with some specific designs having very good reliability. It is worth noting that the bugs–fruit data yielded one of the lowest overall reliabilities, which leads to the rather unusual situation of paradigm's exemplar being less strong than its applications. This may be because other studies assessed categories (e.g., black and white faces) which both belong to a higher-level category (e.g., race). This simplifies the task because it involves a judgment on only one relevant dimension (De Houwer & De Bruycker, 2007). Whatever the reason, the results indicate that topic and stimuli contribute considerably to the reliability of a GNAT. These results give confidence that a well-designed GNAT for measuring a clearly defined construct can have good reliability.

Two block difference scores

It has long been known (e.g., Lord, 1963) that difference scores generally have lower reliabilities than their component scores, despite their surface appearance of a test double the length. The results of Study 1 empirically confirm that two-block GNAT difference scores also have lower reliabilities than their component blocks, and often fail to reach the recommended minimum for research. However, difference score reliability is a more complex issue than it first appears (e.g., Williams & Zimmerman, 1996) and the concept of classical test reliability may not be entirely appropriate for some applications where difference scores are used (e.g., Mellenbergh, 1999). We do not wish to discuss these points in detail, but suggest that the implications of apparent low reliability of difference scores should be considered when interpreting the many GNAT papers that report differences between "bipolar" pairs of blocks (e.g., "good" and "bad", or "self" and "other"). Furthermore, this finding is relevant to the broader debate concerning the dimensionality of implicit attitudes (e.g., unidimensional constructs — see Fazio, 2007; Petty, Briñol, & DeMarree, 2007; versus multidimensional construct — see Wilson et al., 2000).

Implications for block length and GNAT design

At least some of the attenuation in reliability observed for the first half/second half splits is due to learning. Evidence for this conclusion is the marked increase of d' for the second half of each block, while d' scores from odd/even halves did not meaningfully differ. Quite apart from reliability considerations, this finding indicates that longer blocks will generally result in higher d' scores, as a consequence researchers should use equal block lengths for all blocks that will be compared to avoid spurious between-block differences.

Finally, it must be kept in mind that because the block length is effectively halved by the splitting process, the results from Study 1 underestimate the actual reliability of these GNATs. Correcting this bias is the subject of the next section of this paper.

Study 2: the effect of block length on reliability

For two tests of different lengths with equally good items, the longer test will have a higher reliability, therefore, split half correlations underestimate a test's true reliability. This fact led to the development of the Spearman–Brown "prophecy" formula, which estimates the expected reliability of a test if more items of equal quality are added to a test of known reliability. The prophecy formula, like Cronbach's alpha, relies on algebraic properties of test items that do not hold for individual trials in a d' calculation and, thus, cannot be used to correct reliability estimates for the GNAT. Aside from the constructs under investigation, and the items used to represent them, block length is likely to be the biggest determinant of a GNAT's reliability.

A conceptual model for GNAT reliability

We require a model for GNAT reliability that accounts for typical GNAT contextual variables (e.g., practice effects, item sampling). Developing such a model would involve making many assumptions; however, we can avoid this by using a bootstrap-like approach. We begin by obtaining a dataset from a representative sample. We treat this as an approximation of the universe of trials and draw random samples with replacement from this. Each (re)sample is a "virtual" independent trial. We thus empirically approximate the ideal reliability study – an independent, memory-less repetition – the only requirement being that the original sample adequately reflects the behavior of the target population, empirically capturing idiosyncrasies of the distribution of responses and dependencies between trials.

Study 2 aimed to determine how reliability is affected by block length using the proposed resampling method. This information can be used to estimate the number of trials required to reach target reliability for a given set of items. In addition, this will also provide an indication of the extent to which uncorrected split-half procedures underestimate reliability.

Method

Analysis

The data from Study 1 was used. Pairs of GNAT blocks ranging from 20 to 160 items in 10 item increments were created by randomly sampling each dataset with replacement to create a "test" and "retest" block for each participant. d' was computed for each block in a pair, and the correlation between these d' pairs was calculated across participants. One-thousand replications were computed for each dataset at each test length and the resulting distributions were tabulated. This procedure is conceptually similar to the RaSSH of Study 1, except that resampling with replacement allows the creation of blocks or arbitrary length. Since this method produces consistency estimates for the target test length we will call it the Monte Carlo Alpha-Like Coefficient or *MCALC* for short. These operations were completed using scripts written by the authors in the R programming language (version 2.7; R Development Core Team, 2008).

Results

Boxplots of the distributions of correlations were created for each dataset and plotted as a function of block length. These plots can be seen for the attitudes data in Fig. 1. Plots for other datasets are available in an on-line supplement. The overall shape of the reliability-block length function was extremely consistent across all the blocks in each dataset. With the exception of small ripples at block lengths between 100 and 160 trials, the functions were monotonically increasing and negatively accelerating for all GNATs. The mean and median were close to equal, with a tendency for the mean to be slightly lower than the median for the longest blocks. At all block lengths the distributions were quite leptokurtic. With very few (apparently random) exceptions, the distributions of correlations were slightly negatively skewed, although this skew was mostly in the tails outside the quartiles, with the quartiles being fairly symmetrically spaced around the median.

We attempted to characterize the shape of the curves using polynomials and other common functions and found that the Spearman–Brown formula yielded predictions as accurate as any other empirically fitted curve. The discrepancies between empirically obtained MCALC values and those that predicted all possible block doublings (20–40, 30–60, 40–80, 50–100, 60–120, 70–140, 80–160) were very small, with mean prediction errors of only .0072 ($SD = 0.0090$), .0034 ($SD = 0.0156$), .0054 ($SD = 0.0242$) and .0010 ($SD = 0.0122$) for bugs–fruit, implicit racism, implicit ageism and implicit homophobia, respectively. The mean discrepancies for folk psychiatry and familiarity were .0017 ($SD = 0.0061$), and −.0013, ($SD = 0.0167$).

Finally we estimated the block length required to reach reliabilities of .60, .80, and .90 for each study. This was done by interpolating between datapoints on each individual curve, reading the block lengths associated with each target reliability level and averaging these within each study separately. Some designs did not reach reliabilities of .90 for even the longest blocks. The only block in the bugs–fruit design to reach a reliability of .90 was the BUGS–BAD and only at 160 trials. None of the other bugs–fruit blocks reached .90 by this block length. Two of the folk psychiatry blocks failed to reach a reliability of .90, but the remaining blocks did so with 106.67 trials. Only three of the eight[4] blocks in the familiarity design reached .90 and then only at the longest block lengths. None of the ageism, racism or homophobia blocks reached .90, but several came close (.85–.88 at 160 trials). All blocks in all designs except ageism

[4] Note that we did not include the flowers and direction/temperature practice blocks in the calculations for the familiarity study.

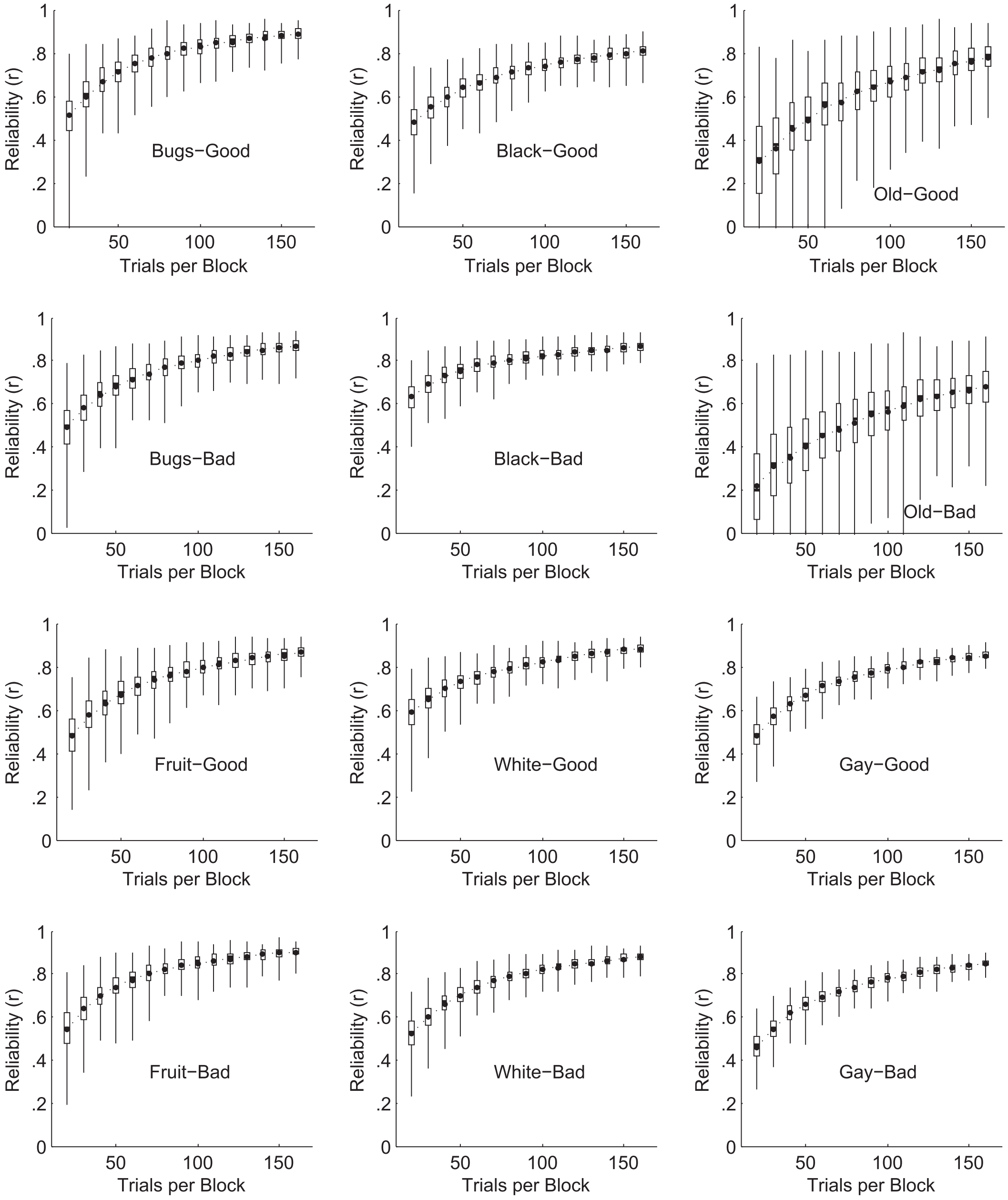

Fig. 1. Distributions of MCALC (resampling) reliability estimates as a function of block length for the bugs-fruit, ageism, racism, and homophobia GNATs (see text). Dots indicate means, the small bars medians, the extremes of the boxes show the upper and lower quartiles, while the whiskers show the maximum and minimum reliability correlations from 1000 replications for each block length. Block lengths range from 20 to 160 trials.

were able to reach reliabilities of .80. The bugs–fruit design reached a reliability of .60 and .80 with an average of 31.75 (*SD* = 4.27) and 88.25 (*SD* = 15.88) trials per block respectively. The required number of trials per block to reach these reliabilities was 19.36 (*SD* = 10.87) and 58.00 (*SD* = 20.48) for folk psychiatry, and 24.50 (*SD* = 6.52) and 75.25 (*SD* = 10.07) for familiarity, 36.00 (*SD* = 1.41) and 112.50 (*SD* = 3.54) for homophobia, and 26.75 (*SD* = 10.75) and 101.25 (*SD* = 32.76) for racism. The ageism design reached .60 with an average of 85.00 trials (*SD* = 21.21).

Discussion

The results show that GNAT reliability increases as a monotonic and negatively accelerating function of block length. Thus, increasing the number of trials brings an ever-diminishing return, and for a set of items there will be an optimum number of trials which balances block length against reliability. The shape of the reliability-block length function was consistent across datasets, but the overall magnitude of the reliability varied. It is clear that the constructs being measured and, potentially the items representing them, affect the reliability of a GNAT. The results also show that initially more reliable item sets gain reliability at a faster rate with increasing block length than initially less reliable item sets.

Are there any conclusions we can draw about which constructs yield the best reliabilities? From the data, it can be seen that reliabilities of .80 are obtained with as few as 60 trials per block if constructs are highly related (e.g., a cluster of mental disorder conditions and the cognitive factor most typically used to describe them). In contrast, more complex (e.g., ambivalent attitudes where the category is implicitly associated with both positive and negative attributes) or abstract constructs (i.e., nonwords) require around 80 trials per block to achieve good reliability. Within a given domain, however, it is not clear whether congruence in itself (and hence ease of association) influences reliability.

Somewhat surprisingly, very few trials seem to be needed to obtain a minimum reliability of .60 (e.g., approximately 30 trials appear to suffice for a well-designed GNAT). There is, however, some variability between blocks in each design and the weakest blocks often required 10–20 extra trials more than the average to match the target reliability. Interestingly, there is no effect of target-attribute congruence. Congruent blocks are not noticeably more reliable. Given the observed variability, researchers need to carefully consider whether the overall or blockwise reliability is important in their designs.

General discussion

What can we say about the reliability of the GNAT in general? It must be emphasized that a test's reliability is a function of its structure, format, and content. It would seem absurd to expect a pencil and paper personality test of randomly generated items to have reliability as good as, say, the NEO-PI-R (Costa & McCrae, 1992). Similarly, we cannot expect a GNAT for a randomly selected construct comprising untested items to have a high reliability simply because it is a GNAT. However, in contrast to the gloomy view of GNAT reliability held by some, the present paper shows that GNATs *can* achieve good reliability and, given the wide range of constructs examined, we feel confident that our results represent likely values of GNATs for other constructs. Based on our results a figure of 30–40 trials per block is indicated as a rough starting point for creating a GNAT with "acceptable" reliability ($r > .60$–$.70$) and 80–90 trials per block is likely to yield very good ($r > .80$) reliability. It seems that reliabilities of .90 are very hard to achieve; the number of trials required is extremely large and likely to be burdensome to participants. A GNAT could be more precisely designed by collecting pilot data with 40–50 trial blocks and then using the MCALC procedure to estimate the appropriate block length for a chosen level of reliability.

The results of Study 2 indicate that the reliability coefficients from split-half estimates (cf., Study 1) underestimate the reliability of full-length blocks. Comparison of results from Study 1 and 2 suggests that split-half estimates should be revised upwards by 10–20% depending on the construct and block length and, while there is no reason that the Spearman–Brown prediction formula should apply to correcting split half reliability estimates of GNAT d' scores, our results indicate that in practice it is a good empirical approximation. However, using pilot data to simulate the reliability of different block lengths, as we have done in Study 2 is preferable and far more defensible. R source code that researchers can use in their own studies is available from the authors.

We have presented a logical argument for why widely used methods (such as the desirable Cronbach's alpha) cannot be directly applied to GNATs, the relative merits of existing methods, and why standard corrections cannot be applied to split half estimates of GNAT reliability. We have advanced a conceptual argument for a statistic which should be a good reliability indicator and have also shown how empirical distributions of reliability estimates can be used to interpret reliability estimates derived from various split-half methods.

Reliability considerations in designing GNATs and diagnosing GNAT problems

Until a more complete solution to the problem of GNAT reliability is discovered we recommend the following approach for designing GNATs and assessing their reliability. Firstly, calculate split-half reliabilities using odd/even and first-half/second-half splits, and obtain the distribution of a large number of random split half reliability estimates (e.g., RaSSH). If odd/even reliability and the RaSSH mean are similar, researchers should have confidence that the reliability of the GNAT is not unduly influenced by practice effects, but must recognize that both statistics underestimate the true reliability. Researchers can then use the MCALC, or at a pinch, the Spearman–Brown formula as a guide to the true reliability (i.e., corrected for test length). Researchers designing new GNATs could use the MCALC on pilot data to estimate the block length required to achieve a given level of reliability. A tight RaSSH or MCALC distribution can be considered evidence against GNAT sensitivity to both sequence and item-sampling effects and is also consistent with (but does not guarantee) that the GNAT is measuring a single construct. We have not yet investigated "rational" GNAT design, calculating reliability statistics for different item combinations to identify poor versus good items. Should the above conditions not hold, GNAT designers can consider the following points in diagnosing GNAT problems:

1. *Investigate carefully instances where the odd/even and RaSSH mean reliability estimates differ significantly.* Since the odd/even split balances practice and fatigue in the two halves, one might expect these halves to show higher correlations than randomly chosen splits. In the GNATs we studied that the odd/even estimate was generally slightly *lower* than the RaSSH mean. Because the RaSSH averages out random fluctuations and item sampling effects, RaSSH means larger than the odd/even correlation most likely indicate that random and item sampling effects have a bigger impact on GNAT score consistency than learning effects. The reverse would be true where the odd/even reliability is greater than the RaSSH mean.
2. *Use the same block length for all blocks that will be compared. Choose block length carefully, particularly when large differences in d' are observed between the first-half and second-half of a block or the first-half/second-half reliability differs markedly from the RaSSH mean.* The d' for the second half of GNAT blocks was generally higher than that of the first in all our designs, indicating learning effects. Large differences between d' for the first and second half of a block or large differences between reliability estimates derived from first/second half splits compared with other methods should alert researchers to a critical dependency of d' magnitude and consistency on block length. These undesirable effects should generally decrease with increasing block length. Although we have never seen designs that use different block lengths, the present study provides a strong case for *always* using equal block length to avoid spurious between-block differences in d', particularly for short blocks.
3. *Examine RaSSH variability for an indication of item quality and sequence effects.* If the RaSSH distribution shows high variability or is platykurtic, block length and item characteristics need to be examined. Even if the odd/even and RaSSH means agree, high RaSSH distribution variability for short blocks (e.g., less than 40 trials) indicates that blocks may be too short to yield stable scores. Wide RaSSH dispersion for long blocks or a platykurtic RaSSH distribution indicates undesirable levels of item-sampling variability and suggests that items are of poor quality, are heterogeneous, or are tapping multiple constructs. Strong local sequence effects could also be indicated by high RaSSH variability coupled with a leptokurtic RaSSH distribution, where a few particular trial combinations yield extreme values. RaSSH distributional anomalies coupled with large discrepancies between different reliability estimators should be considered particularly problematic.

Summary and conclusion

Tests used in social and clinical psychology are required to come with some statement about their reliability, without which they are interpreted with suspicion at best, and simply not used at worst. Our results indicate that GNATs can be reliable and that simple alternating item split-half correlation provides a usable estimate of reliability which can be stabilized against sampling anomalies by comparing it to a distribution of random splits. However, both these procedures generally underestimate reliability. We have provided a method for estimating GNAT reliability that can be viewed as a Monte Carlo analog of Cronbach's alpha or an empirical bootstrap approximation of Guttman's idealized replication. The method uses sampling *with* replacement to simulate test/retest blocks of the appropriate length.

We recognize that our approach has some limitations, namely, that it quantifies only random errors but not systematic errors inherent in an instrument. Our MCALC approach lumps method variance with the true score. Since this is also true of classical indices like Cronbach's alpha, we do not consider this a fatal weakness of this approach so long as this limitation is understood. More detailed generalizability studies are needed to determine what portion of the GNAT score variance is attributable to process and individual variance, and ultimately a model for the cognitive process underlying GNAT responses is desirable.

In the absence of any other guides, we cautiously advance the “magic numbers” of 40 and 80 trials per block as minimums likely to yield “minimally acceptable” and “good” reliability respectively, but urge researchers to determine the actual reliability of their own GNATs from empirical data using the RaSSH or the MCALC. Overall, our results provide good reason to be optimistic about the reliability of the GNAT. The reliability estimates obtained here show that GNAT reliability can be considerably better than previous publications indicate. Given the diverse content of the GNATs we have studied, we are confident that similar levels of reliability would be readily achieved by any well-designed GNAT.

While there is still much to be learned about implicit measures, the present study shows that the GNAT is reliable enough to be a highly useful research tool for social psychology. It can achieve reliabilities comparable to those of its strongest competitor, the IAT. It is possible that a GNAT requires more trials than a comparable IAT to achieve a given level of reliability, however, the generally shorter trial lengths of a GNAT allow more trails to be tested in a given time, and such increases are more than offset by the other advantages the GNAT confers. Researchers should continue using the GNAT with increased confidence, and those searching for an implicit measure of association now have another good reason to consider the GNAT.

Supplementary materials related to this article can be found online at doi:10.1016/j.jesp.2012.03.001.

References

Allen, T. J., Sherman, J. W., Conrey, F. R., & Stroessner, S. J. (2009). Stereotype strength and attentional bias: Preference for confirming versus disconfirming information depends on processing capacity. *Journal of Experimental Social Psychology*, *45*(5), 1081–1087.

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory.* Long Grove, IL: Waveland Press.

Anderson, J., & Kaufmann, L. M. (2011). Exploring attitudes towards homosexuality: The role of attitude relevant factors and context effects. *Paper presented at 9th Biennial conference of the Asian Association of Social Psychology, Kunming, China*.

Bassett, J. F., & Dabbs, J. M., Jr. (2005). A portable version of the go/no go association task (GNAT). *Behavior Research Methods*, *37*(3), 506–512.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*(5), 828–841.

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*, 192–212.

Bluemke, M., & Friese, M. (2006). Do features of stimuli influence IAT effects? *Journal of Experimental Social Psychology*, *42*(2), 163–176.

Boldero, J., Rawlings, D. R., & Haslam, N. (2007). Convergence between GNAT-assessed implicit and explicit personality. *European Journal of Personality*, *21*(3), 341–358.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revised? *Journal of Personality and Social Psychology*, *79*(4), 631–643.

Boucher, H. C., Peng, K., Shi, J., & Wang, L. (2009). Culture and implicit self-esteem: Chinese are ‘good’ and ‘bad’ at the same time. *Journal of Cross-Cultural Psychology*, *40*(1), 24–45.

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology*, *81*(5), 760–777.

Brunel, F. F., Tietje, B. C., & Greenwald, A. G. (2004). Is the Implicit Association Test a valid and valuable measure of implicit consumer social cognition? *Journal of Consumer Psychology*, *14*(4), 385–404.

Conner, T., & Barrett, L. F. (2005). Implicit self-attitudes predict spontaneous affect in daily life. *Emotion*, *5*(4), 476–488.

Costa, P. T., Jr., & McCrae, R. R. (1992). *NEO PI-R professional manual.* Odessa, FL: Psychological Assessment Resources, Inc..

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*, 349–354.

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

Cvencek, D., Greenwald, A. G., Brown, A., Snowden, R., & Gray, N. (2010). Faking of the Implicit Association Test is statistically detectable and partly correctable. *Basic and Applied Psychology*, *32*(4), 302–314.

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

De Houwer, J., & De Bruycker, E. (2007). The Implicit Association Test outperforms the extrinsic affective Simon task as an implicit measure of inter-individual differences in attitudes. *British Journal of Social Psychology*, *46*(2), 401–421.

De Houwer, J., & Moors, A. (2007). How to define and examine the implicitness of implicit measures. In B. Wittenbrink, & N. Schwartz (Eds.), *Implicit measures of attitudes* (pp. 179–194). New York, NY: Guilford Press.

Devos, T., Viera, E., Diaz, P., & Dunn, R. (2007). Influence of motherhood on the implicit academic self-concept of female college students: Distinct effects of subtle exposure to cues and directed thinking. *European Journal of Psychology of Education*, *22*(3), 371–386.

Doherty, K., & Schlenker, B. R. (1991). Self-consciousness and strategic self-presentation. *Journal of Personality*, *59*(1), 1–18.

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, *25*(5), 603–637.

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), 1013–1027.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and uses. *Annual Review of Psychology*, *54*, 297–327.

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the implicit association tests. *Basic and Applied Social Psychology*, *27*(4), 307–316.

Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Boston, MA: Houghton Mifflin.

Gonsalkorale, K., von Hippel, W., Sherman, J. W., & Klauer, K. C. (2009). Bias and regulation of bias in intergroup interactions: Implicit attitudes toward Muslims and interaction quality. *Journal of Experimental Social Psychology*, *45*(1), 161–166.

Green, J. A., & Swets, D. M. (1966). *Signal detection theory and psychophysics.* New York, NY: Wiley.

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. (2001). Health of the Implicit Association Test at age 3. *Zeitschrift für Experimentelle Psychologie*, *48*, 85–93.

Greenwald, A. G., Nosek, B. A., & Sriram, N. (2006). Consequential validity of the Implicit Association Test. *American Psychologist*, *61*(1), 56–61.

Guttman, L. (1945). A basis for analyzing test–retest reliability. *Psychometrika*, *10*(4), 255–282.

Hammond, K. R. (1948). Measuring attitudes by error-choice: An indirect method. *Journal of Abnormal and Social Psychology*, *43*(1), 38–48.

Haslam, N. (2005). Dimension of folk psychiatry. *Review of General Psychology*, *9*(1), 35–47.

Haslam, N., Ban, L., & Kaufmann, L. M. (2007). Lay conception of mental disorder: The folk psychiatry model. *Australian Psychologist*, *42*(2), 129–137.

Inquisit 2.0.60616 [Computer software] (2006). Seattle, WA: Millisecond Software.

Kaufmann, L. M. (2010). Debugging the GNAT. Unpublished PhD thesis. University of Melbourne, Melbourne, Australia.

Kaufmann, L. M., & Haslam, N. (2006). The role of familiarity in implicit evaluations. *Paper presented at the 35th annual conference of the Society for Australasian Social Psychologists, Canberra, Australia*.

Kaufmann, L. M., & Haslam, N. (2007). Implicit associations measuring more than implicit evaluations: A Go/No Go test of the Folk Psychiatry model. *Poster presented at the 2007 conference for the Society of Personality and Social Psychology, Memphis, TN, United States*.

Kaufmann, L. M., Haslam, N. (submitted for publication). *Debugging the GNAT: Assessing the fakeability Go/No Go Association Task*.

Kaufmann, L. M., Johnson, B. (2011). Exploring explicit and implicit colour blind racial attitudes in Australia. Unpublished manuscript.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 21–38). Madison, WI: University of Wisconsin Press.

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

Loughnan, S., & Haslam, N. (2007). Animals and androids: Implicit associations between social categories and nonhumans. *Psychological Science*, *18*(2), 116–121.

McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In J. F. Dovidio, & S. L. Gaertner (Eds.), *Prejudice, discrimination and racism* (pp. 91–126). New York, NY: Academic.

Mellenbergh, G. J. (1999). A note on simple gain score prediction. *Applied Psychological Measurement*, *23*(1), 87–89.

Miller, J. (1996). The sampling distribution of d′. *Perception & Psychophysics*, *58*(1), 65–72.

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluations. *Journal of Experimental Social Psychology*, *132*(3), 455–469.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know. *Psychological Review*, *84*(3), 231–259.

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-Go Association Task. *Social Cognition*, *19*(6), 625–664.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., et al. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*, 36–88.

Nunnally, J. C. (1967). *Psychometric theory.* New York, NY: McGraw-Hill.

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

Petty, R. E., Briñol, P., & DeMarree, K. G. (2007). The Meta-Cognitive Model (MCM) of attitudes: Implications for attitude measurement, change, and strength. *Social Cognition*, *25*(5), 657–686.

R Development Core Team (2008). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing3-900051-07-0 URL. http://www.R-project.org

Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology*, *44*(2), 386–396.

Rudman, L. A., & Heppen, J. B. (2003). Implicit romantic fantasies and women's interest in personal power: A glass slipper effect? *Personality and Social Psychology Bulletin*, *29*(11), 1357–1370.

Rudolph, A., Shröder-Abé, M., Schütz, A., Gregg, A. P., & Sedikides, C. (2008). Through a glass, less darkly? Reassessing convergent and discriminant validity in measures of implicit self-esteem. *European Journal of Psychological Assessment*, *24*(4), 273–281.

Sanderson, C. A., & Cantor, N. (2001). The association of intimacy goals and marital satisfaction: A test of four mediational hypotheses. *Personality and Social Psychology Bulletin*, *27*, 1567–1577.

Schlenker, B. R., Bonoma, T. V., & Tedeschi, J. T. (1973). Impression management revisited. *American Psychologist*, *28*, 360.

Schmukle, S. C., Back, M. D., & Egloff, B. (2008). Validity of the five-factor model for the implicit self-concept of personality. *European Journal of Psychological Assessment*, *24*(4), 263–272.

Schnabel, K., Asendorpf, J. B., & Greenwald, A. G. (2008). Assessment of individual differences in implicit cognition: A review of IAT measures. *European Journal of Psychological Assessment*, *24*(4), 210–217.

Scopus [database] (2012). Amsterdam, Netherlands: Elsevier B.V..

Sherman, J. W., Stroessner, S. J., Conrey, F. R., & Azam, O. A. (2005). Prejudice and stereotype maintenance processes: Attention, attribution and individuation. *Journal of Personality and Social Psychology*, *89*(4), 607–622.

Siegle, G. J. (1994). The balanced affective word list creation program. Available Web:. http://www.sci.sdsu.edu/CAL/wordlist/

Smith, V. J., Stewart, T. L., Myers, A. C., & Latu, I. M. (2008). Implicit coping responses to racism predict African American's level of psychological distress. *Basic and Applied Social Psychology*, *30*(3), 264–277.

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

Spence, A., & Townsend, E. (2006). Implicit attitudes towards genetically modified (GM) foods: A comparison of context-free and context-dependent evaluations. *Appetite*, *46*(1), 67–74.

Spence, A., & Townsend, E. (2007). Predicting behavior towards genetically modified food using implicit and explicit attitudes. *British Journal of Social Psychology*, *46*(2), 437–457.

Steffens, M. C. (2004). Is the Implicit Association Test immune to faking? *Experimental Psychology*, *51*(3), 165–179.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*, 99–103.

Teachman, B. A. (2007). Evaluating implicit spider fear associations using the Go/No Go Association Task. *Journal of Behavior Therapy and Experimental Psychiatry*, *38*(2), 156–167.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications.* Thousand Oaks, CA: Sage.

Wickens, T. D. (2002). *Elementary signal detection theory.* New York, NY: Oxford University Press.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, *20*(1), 59–69.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*(1), 101–126.