

Showing that the race model inequality is not violated

Matthias Gondan · Verena Riehl · Steven P. Blurton

Published online: 20 August 2011
© Psychonomic Society, Inc. 2011

Abstract When participants are asked to respond in the same way to stimuli from different sources (e.g., auditory and visual), responses are often observed to be substantially faster when both stimuli are presented simultaneously (redundancy gain). Different models account for this effect, the two most important being race models and coactivation models. Redundancy gains consistent with the race model have an upper limit, however, which is given by the well-known race model inequality (Miller, 1982). A number of statistical tests have been proposed for testing the race model inequality in single participants and groups of participants. All of these tests use the race model as the null hypothesis, and rejection of the null hypothesis is considered evidence in favor of coactivation. We introduce a statistical test in which the race model prediction is the alternative hypothesis. This test controls the Type I error if a theory predicts that the race model prediction holds in a given experimental condition.

Keywords Divided attention · Cognitive architecture · Non-inferiority · Model testing and validation

Electronic supplementary material The online version of this article (doi:10.3758/s13428-011-0147-z) contains supplementary material, which is available to authorized users.

M. Gondan (✉)
Institute of Medical Biometry and Informatics,
University of Heidelberg,
Im Neuenheimer Feld 305,
69120 Heidelberg, Germany
e-mail: gondan@imbi.uni-heidelberg.de

V. Riehl · S. P. Blurton
Department of Psychology, University of Regensburg,
Regensburg, Germany

Redundant-signals effect

The redundant-signals task might be considered one of the most basic paradigms in research on cognitive architecture. In a typical redundant-signals experiment, participants receive stimuli from two different sources (hereafter, A and V, for auditory and visual stimuli, respectively; of course, the present results generalize to any combination of signals, within or across sensory modalities). The critical aspect is that the same speeded response is required for both A and V—for example, a simple manual response or a given choice. In a third condition, both stimuli are presented simultaneously (redundant signals, AV); in this condition, responses are often observed to be substantially faster than in the single-signal conditions A and V. At first glance, this redundancy gain, in itself, indicates some sort of integration of the information provided by the two signals. However, different models can account for the effect, including serial, parallel, and coactivation models of information processing (e.g., Miller, 1982; Schwarz, 1994; Townsend & Nozawa, 1997).

In the analysis of response times observed in a redundant-signals task, a general distinction is often made between separate activation and coactivation models. The information provided by the different sensory systems might be processed in separate pathways (separate-activation models; e.g., race model, serial self-terminating model), or it might be pooled into a common channel and processed as a combined entity (coactivation). The most important member of the class of separate-activation models is the so-called *race model*: The race model assumes that processing of a redundant AV stimulus occurs in separate channels; the overall processing time D_{AV} is then determined by the faster of the two channels: $D_{AV} = \min(D_A, D_V)$. If the processing-time distribution D_A is invariant

in A and AV, and \mathbf{D}_V is invariant in V and AV (context invariance; see, e.g., Luce, 1986, p. 130), the minimum rule yields, on average, faster processing of AV than of either A or V alone. The redundancy gain according to the race model has an upper limit, however. This upper limit is given by the well-known race model inequality (Miller, 1982),

$$F_{AV}(t) \leq F_A(t) + F_V(t), \text{ for all } t, \quad (1)$$

with $F(t) = P\{\mathbf{T} \leq t\}$ denoting the probability for a response within t milliseconds, and $\mathbf{T} = \mathbf{D} + \mathbf{M}$ denoting the response time, which is usually decomposed into the processing time \mathbf{D} and a context-invariant residual \mathbf{M} (motor execution, finger movement etc.; see Luce, 1986, chap.3). If Inequality (1) holds for all t , the response time distribution for AV is consistent with the race model prediction. Under the race model, the redundancy gain is maximal for $F_{AV}(t) = F_A(t) + F_V(t)$ —more precisely, $F_{AV}(t) = \min[1, F_A(t) + F_V(t)]$, because the left side cannot exceed unity. This maximum is attained in some serial self-terminating models (Appx. B in Gondan, Götze & Greenlee, 2010) and in race models for which context invariance holds and the correlation of the channel-specific processing times \mathbf{D}_A , \mathbf{D}_V is maximally negative (rank correlation -1 ; see, e.g., Colonius, 1990; Townsend & Wenger, 2004).

Violation of Inequality (1) at any t (Fig. 1a) rules out the race model—and, more generally, the entire class of separate activation models (Miller, 1982). Since the race model inequality is based on both separate processing and context invariance, violation of the race model prediction rules out separate activation, or context invariance, or both. For example, race models with mutually facilitating channels have been shown to produce weak violations of the race model inequality (e.g., Mordkoff & Yantis, 1991; Townsend & Wenger, 2004). In most studies, however, a violation of Inequality (1) is interpreted as evidence for integrated processing of the redundant information.

The race model inequality can be generalized in a number of ways—for example, to stimuli presented with onset asynchrony (Miller, 1986), to experiments with catch trials (“kill-the-twin” correction; Eriksen, 1988; Gondan & Heckel, 2008), and to factorial manipulations within the two modalities (Theorem 1 in Townsend & Nozawa, 1995). Here, we focus on an issue related to statistical tests of Inequality (1), but the results hold for these generalizations as well.

As a motivating example, consider a prototypical scenario with two experimental conditions A and B. Theoretical considerations suggest that Inequality (1) is violated in Condition A, whereas it is expected to hold in Condition B (e.g., Feintuch & Cohen, 2002; Schröter, Frei, Ulrich, & Miller, 2009). Feintuch and Cohen presented two

features of a redundant signal either in spatial correspondence (Condition A) or spatially separated (Condition B). In Condition A, the theory predicts coactivation of feature-specific response selectors, whereas in Condition B, redundancy gains were expected to be consistent with separate activation. In tests related to Condition A, the race model takes the role of the null hypothesis. Two types of tests have been developed for this situation, depending on whether Inequality (1) is tested in a single participant (Miller, 1986, pp. 336–337; Maris & Maris, 2003; Vorberg, 2008) or in a group (Gondan, 2010; Miller, 1982; Ulrich, Miller, & Schröter, 2007). We denote these tests as “standard tests” of the race model inequality. These tests demonstrate, at a controlled Type I error, that Inequality (1) is violated at some t . In contrast, for Condition B, the appropriate statistical test has to demonstrate that the observed results are consistent with Inequality (1),

$$\begin{aligned} \text{coac } H_0 : F_{AV}(t) > F_A(t) + F_V(t), \text{ for some } t, \text{ versus} \\ \text{race } H_1 : F_{AV}(t) \leq F_A(t) + F_V(t), \text{ for all } t. \end{aligned} \quad (2)$$

In the inequalities in (2), the race model prediction takes the role of the alternative hypothesis. It is well known that standard significance tests cannot be used to “prove” the null hypothesis. In other words, P values greater than 5% resulting from standard tests of the race model inequality (e.g., Gondan, 2010; Miller, 1986) do not demonstrate that $F_{AV}(t) \leq F_A(t) + F_V(t)$ holds for all t .

Here, we describe a significance test that should be used when theoretical considerations predict that the race model inequality holds (i.e., Condition B). The proposed test controls the Type I error rate if the race model does not hold. In the alternative, the test is consistent—that is, the power increases with sample size when the race model holds. However, as null hypotheses with strict inequalities (2) are difficult to test within the classical null hypothesis testing framework, a so-called *noninferiority margin* needs to be introduced.

Noninferiority tests

Noninferiority tests are members of a more general class of equivalence tests. In applied disciplines, these tests are recommended if the study is designed to establish similarity between two groups or experimental conditions (see D’Agostino, Massaro, & Sullivan, 2003, for an overview). In psychology, only a few claims have been made in favor of equivalence tests—for example, to demonstrate that two therapeutic techniques have similar effects (e.g., Rogers, Howard, & Vessey, 1993; Seaman & Serlin, 1998; Tryon, 2001). For a noninferiority test, a margin $\delta > 0$ is specified, which denotes a small effect in the wrong direction that one

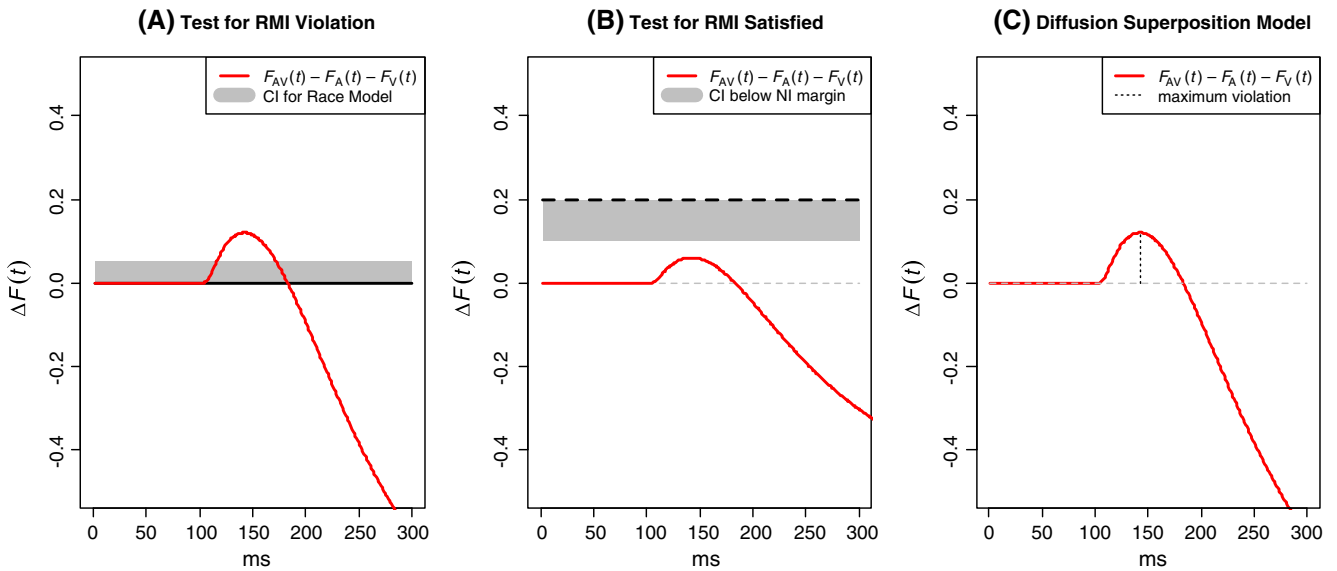


Fig. 1 Tests of the race model. **(a)** The distribution function for AV is greater than the summed distributions for A and V, violating the race model inequality. The significance test is used to demonstrate that $F_{AV}(t) > F_A(t) + F_V(t)$, or, equivalently, that $F_{AV}(t) - F_A(t) - F_V(t)$ is significantly greater than zero, for some t . The direction of the significance test is illustrated by confidence-interval-like gray bars (CI). **(b)** Showing that the race model inequality holds: A significance

test is used to demonstrate that violations are negligible—that is, significantly below the noninferiority margin (NI), $F_{AV}(t) - F_A(t) - F_V(t) < \delta$, for all t . **(c)** Numeric estimate for violation of the race model inequality predicted by the diffusion superposition model (Schwarz, 1994). This estimate can be used for the definition of a noninferiority margin (see the Discussion in the main text)

is willing to tolerate when deciding for the alternative hypothesis. The test is then used for demonstrating non-inferiority; that is, the observed difference is significantly below this margin.

Freitag, Lange, and Munk (2006) proposed a nonparametric noninferiority test for comparison of two distributions: Denote the population distribution functions by $G_1(t)$, $G_2(t)$, with their respective sample distributions $\hat{G}_1(t)$, $\hat{G}_2(t)$. The test is used to demonstrate that $G_1(t)$ stochastically dominates $G_2(t)$ —that is, $G_1(t) \leq G_2(t)$, for all t . Restated in terms of a noninferiority test, $G_1(t) < G_2(t) + \delta$, for all t . The vertical difference $G_1(t) - G_2(t)$ should, thus, never reach or exceed δ ,

$$G_1(t) - G_2(t) < \delta, \text{ for all } t. \tag{3}$$

An intuitive test for the above hypothesis can be constructed using point-wise one-sided confidence intervals for the difference of the sample distributions $\hat{G}_1(t) - \hat{G}_2(t)$. If the upper 95% limit of this confidence interval does not include δ , for all t , violations of stochastic dominance are significantly below the noninferiority margin.

The point-wise approach is, of course, very conservative (see, e.g., Table 1 in Freitag et al., 2006). Bootstrapping can be used to improve the power of the test: If $G_1(t) - G_2(t)$ is everywhere below δ , the maximum of this difference is below δ , as well:

$$\max_t [G_1(t) - G_2(t)] < \delta.$$

This can be shown again statistically using a one-sided confidence interval for the maximum of the vertical distance between the two observed distributions, $d_{\max} = \max_t [\hat{G}_1(t) - \hat{G}_2(t)]$. The confidence interval for this maximum can be determined using so-called *hybrid bootstrapping* (Eq. 4 in Freitag et al., 2006). If the upper 95% limit of the confidence interval around d_{\max} is below δ , noninferiority is established over the entire range of t . Compared to the point-wise test in Inequality (3), the bootstrap distribution of d_{\max} preserves the positive correlation of consecutive values of $\hat{G}(t)$, which substantially increases statistical power (Table 1 in Freitag et al., 2006).

Application to the test of the race model inequality

In order to apply Freitag et al.’s (2006) test to data from a redundant-signals task, an appropriate noninferiority margin must be specified, say $\delta = 0.1$. The hypotheses in (2) are then restated using the noninferiority margin:

$$\begin{aligned} \text{coac } H_0^\delta &: F_{AV}(t) \geq F_A(t) + F_V(t) + \delta, \text{ for some } t, \text{ versus} \\ \text{race } H_1^\delta &: F_{AV}(t) < F_A(t) + F_V(t) + \delta, \text{ for all } t. \end{aligned} \tag{4}$$

In the reformulation of the problem in (4), the non-inferiority margin is defined in probability units (“vertical” test; a horizontal test with the noninferiority margin defined

Table 1 Simulated power for different sample sizes N and non-inferiority margins δ

Test	$N = 30$	$N = 50$	$N = 100$	$N = 200$	$N = 400$
(A) Independent racers					
$\delta = .02$.12	.12	.15	.19	.31
$\delta = .05$.12	.26	.32	.50	.71
$\delta = .10$.33	.44	.63	.84	.97
Classical	.02	.02	.01	.00	.00
(B) Race model with maximum redundancy gain					
$\delta = .02$.04	.04	.05	.06	.11
$\delta = .05$.05	.11	.13	.22	.37
$\delta = .10$.15	.25	.37	.58	.83
Classical	.07	.05	.05	.05	.06
(C) Superposition model					
$\delta = .02$.02	.01	.00	.00	.00
$\delta = .05$.02	.02	.01	.00	.00
$\delta = .10$.05	.06	.04	.04	.02
Classical	.13	.18	.33	.51	.83

Proportions of significant results ($P \leq .05$, 1,000 simulations) for vertical noninferiority tests of the race model inequality with noninferiority margins δ . Classical: proportions of $P > .05$ when testing for violations of the race model inequality (Miller, 1986).

on the milliseconds scale is outlined in Appx. A). If the violation of the race model inequality does not exceed δ , for all t , the null hypothesis in (4) is rejected (Fig. 1b).

Whereas $F_{AV}(t)$ never exceeds unity, the right-hand side must be transformed into a proper distribution function, $F_{AV}(t) \leq \min[1, F_A(t) + F_V(t)]$. The shape of $\min[1, F_A(t) + F_V(t)]$ corresponds to the shape of the lower half of the 1 : 1 mixture of F_A and F_V (Maris & Maris, 2003). A one-tailed $1 - \alpha$ confidence interval for the observed maximum violation of the race model inequality is then built using Freitag et al.'s (2006) algorithm. If the upper limit of the confidence interval is greater than δ , the coactivation H_0 in (4) is retained—namely, that violations of the race model inequality are greater than or equal to δ . If the upper limit of the confidence interval is below δ , violations of the race model inequality are significantly below the noninferiority margin, which favors the race model H_1 (Fig. 1b). The noninferiority test then demonstrates that for a given participant, the observed distribution on the left-hand side of the race model inequality does not substantially exceed the summed distributions of the right-hand side. For this decision, the Type I error is controlled.

Type I error and power

Assuming a specific distribution for the response times in conditions A and V, and assuming that the race model

holds, simulations can be used to generate samples of a specific size. For a given noninferiority margin, it is then possible to estimate the power of the test—that is, the probability that the noninferiority test actually detects that the race model holds, at some prespecified significance level. This power estimate can be used for planning the number of trials required in an experiment. Rejection rates for different sample sizes and noninferiority margins are shown in Table 1 for three scenarios: In the first scenario (Table 1a), a race of stochastically independent (e.g., Eq. 1 in Miller, 1982) channels is assumed, with $F_{AV}(t) = F_A(t) + F_V(t) - F_A(t)F_V(t)$. Response times were generated for condition V67A measured in participant B.D. in Miller (1986). V67A means that AV stimuli were not presented synchronously, but with an onset asynchrony of 67 ms; in this condition, the response time distributions for A and V overlapped maximally. In the second scenario (Table 1b), a separate-activation model with maximum redundancy gain was chosen, $F_{AV}(t) = \min[1, F_A(t) + F_V(t)]$. Not surprisingly, the standard test of the race model inequality (Miller, 1986) rejects the race model in about 5% (i.e., α) of the simulations of Table 1b, mostly independent of the sample size. In contrast, the noninferiority test is consistent, with a positive relationship between power and sample size. However, especially for strict noninferiority margins, the power to detect that the race model inequality holds is rather low, with lowest power at the boundary of the maximally possible redundancy gain. Hence, if the experiment is designed to demonstrate that the race model inequality holds in a given experimental condition, 400 trials per condition or more should be considered (see, e.g., Miller, 1986).

In a third scenario, response times for the same conditions were generated assuming a coactivation model that assumes linear superposition of channel-specific diffusion processes (Schwarz, 1994). Using the parameters of Table 1 in Schwarz (1994, $\rho_{DM} = 0$), the superposition model predicts a substantial violation of the race model inequality (see also Appx. A in Gondan et al., 2010). Table 1c shows that for this setting, the noninferiority keeps the nominal significance level, while the power of a standard test of the race model inequality (Miller, 1986) increases with sample size.

Discussion

Standard procedures for testing the race model use the race model prediction as the null hypothesis (Gondan, 2010; Maris & Maris, 2003; Miller, 1986; Ulrich et al., 2007; Vorberg, 2008). These tests control the Type I error if a theory predicts that the race model inequality is violated in a given experimental condition. In other experimental

conditions, however, theoretical considerations might predict that the race model holds (e.g., Feintuch & Cohen, 2002; Schröter et al., 2009). Standard tests of the race model inequality would then be expected to yield a nonsignificant result. A nonsignificant violation of the race model prediction should, however, not be taken as support for the race model (e.g., Corballis, Hamm, Barnett, & Corballis, 2002; Feintuch & Cohen, 2002, Exp. 1; Grice, Canham, & Gwynne, 1984, Exp. 4). More generally, a nonsignificant test result should not be considered as evidence for the null hypothesis (Altman & Bland, 1995). Classical null hypothesis tests are consistent in the alternative hypothesis only (e.g., Rouder, Speckman, Sun, Morey, & Iverson, 2009): If the alternative hypothesis holds, the power to reject the null hypothesis increases with the precision of the measurement (e.g., sample size). Stated differently, standard tests of the race model inequality can be biased in favor of the race model by collecting small and noisy sets of data, thereby reducing the power of the test to detect violations of the race model prediction.

For such situations, we propose an alternative testing procedure in which the race model prediction takes the role of the alternative hypothesis. The test can be applied to data from single participants (for multiple participants, see Appx. B). For construction of the appropriate test, it was necessary to restate the nonstrict alternative hypothesis [Ineq. (2), ${}^{\text{race}}H_1$] as its strict counterpart [Ineq. (3), ${}^{\text{race}}H_1^\delta$], thereby introducing a noninferiority margin δ . It can then be tested whether violations of the race model inequality are significantly below the noninferiority margin. Powerful tests for noninferiority and stochastic dominance have been proposed by Freitag et al. (2006) and Davidson and Duclos (2009). The following technical aspect of these tests should be pointed out: In the classical comparison of two distributions (e.g., the one-tailed Kolmogorov–Smirnov test; see also the above references to standard tests of the race model inequality), the null hypothesis states that stochastic dominance holds, $G_1(t) \leq G_2(t)$ for all t . The alternative hypothesis assumes that stochastic dominance does not hold, $G_1(t) > G_2(t)$ for *some* t . In contrast, in Freitag et al. and Davidson and Duclos, the alternative hypothesis establishes stochastic dominance—that is, $G_1(t) < G_2(t)$ for *all* t . It is this important property that enables testing the race model prediction over its entire range, at a controlled Type I error probability. Such an alternative hypothesis is much more informative than the alternative hypothesis of the standard test of the race model, but comes at a cost: The new tests are rather conservative, and large samples are needed to obtain reasonable power.

In applied disciplines—in particular, biomedical research—the idea and the principle of equivalence testing and noninferiority testing is certainly not new (e.g., Blackwelder, 1982). Equivalence tests and non-

inferiority tests are now considered standard techniques for showing the similarity of two therapeutic arms (e.g., Allen & Seaman, 2006; Food and Drug Administration, 2001; Wellek, 2003; Westlake, 1988). For the assessment of formal models that are more complex than simple two-group comparisons, equivalence tests and noninferiority tests have only rarely been employed (for a forest growth model, see, e.g., Robinson & Froese, 2004). Rather, Bayesian model comparisons have been suggested for the choice between multiple model candidates (Gallistel, 2009; Wagenmakers, 2007; Wagenmakers, Lee, Lodewyckx, & Iverson, 2008). For the present application, in which only a single model prediction is under consideration, researchers are often satisfied to show that the results are consistent with the prediction, as reflected by a nonsignificant discrepancy measure—for example, a nonsignificant goodness-of-fit statistic, or $P > .05$ in the test of the race model inequality. In these situations, equivalence tests and noninferiority tests seem more appropriate, because they switch the roles of the null and the alternative hypotheses, and consistency of the theory and data is supported by statistical significance.

The new tests have a drawback, however: It is clear that the margin δ must be strictly positive—otherwise, data generated by a race model with $F_{AV}(t) = F_A(t) + F_V(t)$ would not belong to the alternative hypothesis in (4). As a consequence, the model prediction and observed data cannot be said to fit exactly anymore. Rather, close correspondence of the model and data is accepted. In the test outlined in (4), the race model is said to hold if deviations between the observed $\hat{F}_{AV}(t)$ and $\hat{F}_A(t) + \hat{F}_V(t)$ are *significantly below* the noninferiority margin. This conclusion should be given emphasis, because it might be considered one of the main limitations of the proposed new test: Whereas it would be desirable to have a test that indicates that the race model inequality “holds exactly,” the proposed test only states that violations of the model prediction are significantly below the tolerance defined by δ . Stated differently, and perhaps more optimistically, the noninferiority test takes into account that the predictions of abstract models and observed data can rarely be expected to fit exactly. Given a reasonable noninferiority margin, the test rather provides a means to determine if the model describes the data “well enough” (e.g., Serlin & Lapsley, 1993). The latter is again closely related to the definition of the noninferiority margin: If the experimenter desires to test the model prediction with high precision, a small margin is chosen, which in turn requires large samples (Table 1). If a more liberal margin is chosen, precision is lower, and the required sample size will decrease (in the extreme case, $\delta \geq 1$, noninferiority trivially holds).

Of course, the noninferiority margin should be defined before actually running the experiment. In applied disci-

plines, “biocreep” is avoided by choosing a noninferiority margin that reflects a therapeutically irrelevant effect in comparison to the best treatment available (D’Agostino et al., 2003). In contrast, in fundamental research, it would be difficult to justify a specific margin, and for a test of the race model inequality, any choice of δ will be somehow arbitrary.¹ A reasonable size for the noninferiority margin can be determined from parametric coactivation models—for example, the diffusion superposition model (Schwarz, 1994; Townsend & Wenger, 2004, Figs. 14 & 15). For example, with $\mu_A = 1.34$, $\sigma_A = 11.7$, $\mu_V = 0.53$, $\sigma_V = 4.3$, and a criterion fixed at 100 (Table 1 in Schwarz, 1994), the predicted detection times violate Inequality (1) by a maximum amount of 12% (Fig. 1c). This percentage might be too optimistic, because the observed response times include more than just stimulus detection (e.g., response execution, finger movement; see, e.g., the discussion in Schwarz, 1994). The violation predicted by specific coactivation models can, however, serve as a starting point for the definition of a realistic noninferiority margin for the experiment.

In our motivating example, coactivation was expected for Condition A, while separate activation was expected for Condition B. Thus, strictly speaking, both condition-specific predictions should be confirmed in order to support the theory. It is, of course, possible to directly compare the sizes of the violation observed in the two conditions (e.g., by a confidence interval approach; see Miller, 1986, p. 337, right column). Results would be considered consistent with the theory if the size of the race model violation observed in A were significantly higher than the size of the race model violation observed in B. Although this test does not demonstrate that the race model holds in Condition B, the violation observed in Condition A might serve to determine a reasonable limit for the noninferiority margin, as well. In any case, the specific choice of δ should be made transparent to the reader.

We have outlined a test that can be used if the race model is predicted to hold in a given experimental condition. Of course, the test cannot “confirm” the race model in a strict sense. In general, it is not possible to conclude that a given model is correct, just because a single prediction of the model holds in a given set of data. Another, completely different architecture might make the same prediction. More specifically, just because the race model inequality is not violated in a given set of response times, one cannot unambiguously conclude that participants actually processed the information in parallel (see, e.g., Table 1 in Ulrich & Miller, 1997). Our method, thus, cannot overcome

the general limitations of abstract model testing, especially in one-way situations in which only a single model prediction is to be tested. However, we propose a valid statistical procedure to investigate whether this prediction of the model is met by the results. Although the decision of the test depends, by design, on the specific choice of the noninferiority margin, we think it is preferable to use an appropriate statistical test that adequately controls the Type I error and is consistent in the hypothesis of interest, instead of relying on the nonsignificant P value of an inappropriate test.

Author note This research was supported by the German Research Foundation (DFG, GO 1855/1-1 and GR 988/20-2). The authors thank Haiyuan Yang and James Townsend for very helpful comments on our manuscript. The online [supplemental materials](#) of this article include a script (written in R statistical language, R Development Core Team, 2011) that can readily be used for testing whether the race model prediction holds in a given data set, and which may be downloaded from www.springerlink.com.

Appendix A: Showing that the race model inequality horizontally holds

Inequality (1) states that $F_{AV}(t)$ is below $F_A(t) + F_V(t)$, for all t ; the inequality can, however, be read horizontally, as well: The race model predicts that $F_{AV}(t)$ is to the right of $F_A(t) + F_V(t)$ over its entire range. Thus, there is an alternative way to specify the noninferiority test,

$$\begin{aligned} \text{coac} H_0^\delta : F_{AV}(t - \varepsilon) &\geq F_A(t) + F_V(t) + \delta, \text{ for some } t, \text{ versus} \\ \text{race} H_1^\delta : F_{AV}(t - \varepsilon) &< F_A(t) + F_V(t) + \delta, \text{ for all } t. \end{aligned} \quad (\text{A1})$$

with the noninferiority margin ε now being defined in the horizontal direction. Davidson and Duclos (2009) proposed a likelihood ratio test for demonstrating restricted stochastic dominance—that is, $G_1(t) < G_2(t)$, for some interval $t_1 \leq t \leq t_2$. In a first step, the test determines the maximum of the empirical likelihood (Owen, 2001) under the alternative hypothesis—namely, strict stochastic dominance for $t_1 \leq t \leq t_2$. If there is dominance in the two-sample distributions $\hat{G}_1(t)$ and $\hat{G}_2(t)$, the constraint is satisfied. Maximization is then unconstrained, and the maximum only depends on the sample size: $\ell_{\max}(\cdot) = -N_1 \log N_1 - N_2 \log N_2$ (Eq. 4 in Davidson & Duclos, 2009). In a second step, the empirical likelihood is maximized subject to the constraint that the two distributions touch, $G_1(t) = G_2(t)$, for some $t_1 \leq t \leq t_2$ (i.e., the null hypothesis). This constrained maximum $\ell_{\max}(\cdot|t)$ can again be determined analytically (Eq. 12 in Davidson & Duclos, 2009). It can be shown that if the null hypothesis holds, twice the log likelihood ratio of the unconstrained and the constrained maximum $\ell_{\max}(\cdot|H_0) = \max_t \ell_{\max}(\cdot|t)$ asymptotically follows a χ_1^2 distribution. If

¹ A similar argument holds, of course, for the significance level. Why should the consumer risk be exactly $\alpha = 5\%$ when testing a specific model prediction?

the null hypothesis is rejected, stochastic dominance is established for the interval of interest (t_1, t_2) . It should be underlined that \mathcal{L}_{\max} corresponds again to a point-wise test for all possible touch points within the interval of interest. The test is, therefore, very conservative. Davidson and Duclos (Theorem 3) described a weighted bootstrap procedure that can be used to increase the power of the test.

The test proposed by Davidson and Duclos (2009) can easily be generalized to the horizontal noninferiority test of the race model inequality outlined in (A1): Having specified a suitable noninferiority margin ($\varepsilon = 10$ ms, say), the response time distribution for AV is shifted ε milliseconds to the right. The right-hand side of the race model inequality corresponds again to the lower 50% of the mixture of the response times for unimodal stimuli. If the test indicates that the shifted AV distribution dominates the sum of the unimodal distributions in some interval of interest (e.g., 150 . . . 300 ms), the race model inequality has been shown to hold at a controlled Type I error.

A critical aspect of Definition (A1) is that the noninferiority margin ε is then expressed in milliseconds and, therefore, corresponds to the original scale of the response time measurement. Whereas this might be considered an advantage over the vertical definition in (4), the horizontal test depends on the overall speed of the participant. Consider two participants, with Participant 1 responding twice as fast as Participant 2 in all experimental conditions. In this case, the horizontal distance between $F_{AV}(t)$ and $F_A(t) + F_V(t)$ shrinks by a factor of 2. This affects the result of the horizontal noninferiority test; if the noninferiority margin is unchanged, it is well possible that the response times of Participant 2 are in line with the race model prediction, whereas the response times of Participant 1 are not. The vertical test described in the main text uses only the ranks of the data; it is, thus, invariant with respect to order-preserving transformations of the time scale (e.g., logarithm of t). For this reason, we recommend using the vertical noninferiority test.

Appendix B: Multiple participants

The noninferiority test proposed here requires a substantial number of trials (see Table 1) and can be applied to data from single participants only (cf. Maris & Maris, 2003; Miller, 1986; Vorberg, 2008). It is not straightforward to combine results across multiple participants. Strictly speaking, the race model inequality holds for a group of M participants if it holds in each single participant. This argument suggests an intersection–union approach (e.g., Berger, 1982) in which the alternative hypothesis for the group is accepted if and only if all

participant-specific alternative hypotheses are accepted. (Conversely, if the theory predicts that the race model inequality is violated, it would, in principle, suffice to demonstrate that it is violated in a single participant or in the “average participant”; see, e.g., Gondan, 2010; Miller, 1982; Ulrich et al., 2007).

Of course, showing that the race model prediction holds in every participant would be a very conservative approach, and the global test would not be consistent, because statistical power to detect the alternative would decline with increasing M . A less strict approach would be based on the response time distributions of the “average participant”: As in Miller (1982; see also Ulrich et al., 2007), percentiles differences between $F_{AV}(t)$ and $F_A(t) + F_V(t)$ are determined for each participant i , $d_{pi} = \widehat{F}_{AV}^{-1}(p) [\widehat{F}_A + \widehat{F}_V]^{-1}(p)$, $p = 5\%, 10\%, \dots, 50\%$. If the race model prediction holds, $d_{pi} \leq 0$, and so is the average $d_p \leq 0$, for all p . This prediction could be tested using percentile-specific t -tests with a horizontal noninferiority margin $\varepsilon > 0$: If t -tests indicate that d_p is significantly below ε for all percentiles p , the race model inequality is satisfied for the average participant. This test is consistent in both trials and participants.

References

- Allen, I. E., & Seaman, C. A. (2006). Different, equivalent or both? *Quality Progress*, 39, 77.
- Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal*, 311, 485.
- Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24, 295–300.
- Blackwelder, W. C. (1982). “Proving the null hypothesis” in clinical trials. *Controlled Clinical Trials*, 3, 345–353.
- Colonius, H. (1990). Possibly dependent probability summation of reaction time. *Journal of Mathematical Psychology*, 34, 253–275. doi:10.1016/0022-2496(90)90032-5
- Corballis, M. C., Hamm, J. P., Barnett, K. J., & Corballis, P. M. (2002). Paradoxical interhemispheric summation in the split brain. *Journal of Cognitive Neuroscience*, 14, 1151–1157. doi:10.1162/089892902760807168
- D’Agostino, R. B., Sr., Massaro, J. M., & Sullivan, L. M. (2003). Non-inferiority trials: Design concepts and issues—The encounters of academic consultants in statistics. *Statistics in Medicine*, 22, 169–186.
- Davidson, R., & Duclos, J.Y. (2009). *Testing for restricted stochastic dominance* (Document de travail 2009-39, GREQAM). Available at <http://halshs.archives-ouvertes.fr/docs/00/44/35/60/PDF/DT2009-39.pdf>
- Eriksen, C. W. (1988). A source of error in attempts to distinguish coactivation from separate activation in the perception of redundant targets. *Perception & Psychophysics*, 44, 191–193. doi:10.3758/BF03208712
- Feintuch, U., & Cohen, A. (2002). Visual attention and coactivation of response decisions for features from different dimensions. *Psychological Science*, 13, 361–369. doi:10.1111/j.0956-7976.2002.00465.x

- Food and Drug Administration. (2001). *Guidance for industry: Statistical approaches to establishing bioequivalence*. Rockville: Center for Drug Evaluation and Research.
- Freitag, G., Lange, S., & Munk, A. (2006). Nonparametric assessment of non-inferiority with censored data. *Statistics in Medicine*, *25*, 1201–1217.
- Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453. doi:10.1037/a0015251
- Gondan, M. (2010). A permutation test for the race model inequality. *Behavior Research Methods*, *42*, 23–28. doi:10.3758/BRM.42.1.23
- Gondan, M., & Heckel, A. (2008). Testing the race inequality: A simple correction procedure for fast guesses. *Journal of Mathematical Psychology*, *52*, 322–325. doi:10.1016/j.jmp.2008.08.002
- Gondan, M., Götze, C., & Greenlee, M. W. (2010). Redundancy gains in simple responses and go/no-go tasks. *Attention, Perception, & Psychophysics*, *72*, 1692–1709. doi:10.3758/APP.72.6.1692
- Grice, G. R., Canham, L., & Gwynne, J. W. (1984). Absence of a redundant-signals effect in a reaction time task with divided attention. *Perception & Psychophysics*, *36*, 565–570. doi:10.3758/BF03207517
- Luce, R. D. (1986). *Response times: Their role in inferring mental organization*. New York: Oxford University Press, Clarendon Press.
- Maris, G., & Maris, E. (2003). Testing the race model inequality: A nonparametric approach. *Journal of Mathematical Psychology*, *47*, 507–514. doi:10.1016/S0022-2496(03)00062-2
- Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology*, *14*, 247–279. doi:10.1016/0010-0285(82)90010-X
- Miller, J. (1986). Timecourse of coactivation in bimodal divided attention. *Perception & Psychophysics*, *40*, 331–343. doi:10.3758/BF03203025
- Mordkoff, J. T., & Yantis, S. (1991). An interactive race model of divided attention. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 520–538. doi:10.1037/0096-1523.17.2.520
- Owen, A. B. (2001). *Empirical likelihood*. Boca Raton: Chapman & Hall.
- R Development Core Team. (2011). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available from www.R-project.org
- Robinson, A. P., & Froese, R. E. (2004). Model validation using equivalence tests. *Ecological Modeling*, *176*, 349–358.
- Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*, 553–565. doi:10.1037/0033-2909.113.3.553
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. doi:10.3758/PBR.16.2.225
- Schröter, H., Frei, L. S., Ulrich, R., & Miller, J. (2009). The auditory redundant signals effect: An influence of number of stimuli or number of percepts? *Attention, Perception, & Psychophysics*, *71*, 1375–1384. doi:10.3758/APP.71.6.1375
- Schwarz, W. (1994). Diffusion, superposition, and the redundant-targets effect. *Journal of Mathematical Psychology*, *38*, 504–520. doi:10.1006/jmps.1994.1036
- Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods*, *3*, 403–411. doi:10.1037/1082-989X.3.4.403
- Serlin, R. C., & Lapsley, D. K. (1993). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*, 73–83.
- Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, *39*, 321–359. doi:10.1006/jmps.1995.1033
- Townsend, J. T., & Nozawa, G. (1997). Serial exhaustive models can violate the race model inequality: Implications for architecture and capacity. *Psychological Review*, *104*, 595–602. doi:10.1037/0033-295X.104.3.595
- Townsend, J. T., & Wenger, M. J. (2004). A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review*, *111*, 1003–1035. doi:10.1037/0033-295X.111.4.1003
- Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371–386. doi:10.1037/1082-989X.6.4.371
- Ulrich, R., & Miller, J. (1997). Tests of race models for reaction time in experiments with asynchronous redundant signals. *Journal of Mathematical Psychology*, *41*, 367–381. doi:10.1006/jmps.1997.1181
- Ulrich, R., Miller, J., & Schröter, H. (2007). Testing the race model inequality: An algorithm and computer program. *Behavior Research Methods*, *39*, 291–302. doi:10.3758/BF03193160
- Vorberg, D. (2008). *Exact statistical tests of the race model and related inequalities*. Oral presentation at the European Mathematical Psychology Meeting, Graz, Austria
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804. doi:10.3758/BF03194105
- Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York: Springer.
- Wellek, S. (2003). *Testing statistical hypotheses of equivalence*. Boca Raton: Chapman & Hall.
- Westlake, W. J. (1988). Bioavailability and bioequivalence of pharmaceutical formulations. In K. E. Peace (Ed.), *Biopharmaceutical statistics for drug development* (pp. 329–352). New York: Marcel Dekker.