Submitted to the Annals of Statistics

IMPROVED STEPDOWN METHODS FOR ASYMPTOTIC CONTROL OF GENERALIZED ERROR RATES

By Gray Calhoun*

Iowa State University

This paper proposes new stepdown methods for testing multiple hypotheses and constructing confidence intervals while controlling the familywise error rate and other generalized error rates. One method is a refinement of Romano and Wolf's StepM (2005, *Econometrica*) that also removes inequalities that fall outside any $n^{-1/2}$-neighborhood of binding; it has the advantage that the threshold construction is incorporated into the stepdown procedure so it accounts for the number of total hypotheses (leading to better size control than some alternative methods) and excludes more nonbinding inequalities (leading to higher power). This method can also be used to test multiple inequality hypotheses simultaneously and construct confidence intervals for partially identified parameters. The paper presents methods for controlling the k-familywise error rate and the False Discovery Proportion for families of one and two-sided hypotheses as well. The paper also provides Monte Carlo evidence that the methods perform well in finite samples.

1. Introduction. This paper develops improvements to sequential procedures for testing multiple hypotheses. Many existing procedures can lose power when some of the individual null hypotheses hold with parameter values that are far from the alternative—the multiplicity correction is then unnecessarily large and decreases the procedure's power to reject other, false, hypotheses. The canonical example of this problem is testing many one-sided hypotheses (see Hansen, 2005, as well as Andrews, 2012, and Hirano and Porter, 2012, for recent assessments of these issues), but we also present settings where this issue arises for two-sided hypotheses (Theorem 2 and Corollary 2), and our general strategy applies more broadly.

For concreteness, suppose that each hypothesis s is of the form $\theta_s \in \Theta_0$ vs. $\theta_s \in \Theta_a$ and has corresponding test statistic T_s; T_s rejects if it is above some critical value q and the procedure controls the *familywise error rate*

*I would like to thank Helle Bunzel, Brent Kreider, Joseph Romano and Michael Wolf for helpful comments on early versions of this paper.

MSC 2010 subject classifications: Primary 62J15; secondary 62G10

Keywords and phrases: Multiple testing, bootstrap, familywise error rate, false discovery proportion, partial identification, moment inequalities

(FWE) at level α if

$$\Pr[T_s > q \text{ for at least one } s \text{ such that } \theta_s \in \Theta_0] \leq \alpha \tag{1}$$

(we focus on the FWE in this example for simplicity but will present results for other error rates later in the paper; see Section 3). FWE control is stronger than control of the size of the composite hypothesis $\theta_s \in \Theta_0$ for all s, since it must hold for any arrangement of θ_s. This stronger concept is essential if a researcher wants to interpret individual rejections ($T_s > q$) as evidence against the individual hypotheses ($\theta_s \in \Theta_0$).

"Single-step" procedures construct q_1 to control FWE at α and reject all of the individual hypotheses with $T_s > q_1$. A sequential procedure (as in Goeman and Solari, 2010) continues from there by constructing a second critical value q_2 to control FWE at α for the family of hypotheses left after the first step, $\{s : T_s \leq q_1\}$, and rejects all of the hypotheses with $T_s > q_2$. The sequential procedure then continues in the same way, constructing each critical value to control FWE over the remaining hypotheses, until it stops rejecting at, say, the jth step. Somewhat surprisingly, using $q = q_j$ in (1) typically controls the FWE at α (subject to restrictions on the test procedure, of course; see Romano and Wolf, 2005a,b, and Goeman and Solari, 2010, among others).

This paper's method works by identifying regions $\Theta' \subset \Theta_0$ such that the probability that T_s rejects if $\theta \in \Theta'$ is negligible. Instead of testing $\theta_s \in \Theta_0$ against $\theta_s \in \Theta_a$ at level α in each step, we also simultaneously test $\theta_s \in \Theta_0 \setminus \Theta'$ vs. $\theta_s \in \Theta'$ at level ϵ (an arbitrarily small positive quantity). We remove hypothesis s if either test rejects and then proceed sequentially. As ϵ converges to zero, the FWE of this procedure converges to α. Of course, at the end we only reject those hypotheses that were determined to be in Θ_a (i.e. T_s is greater than the last q_j), but removing the additional hypotheses in Θ' during the sequential process can increase the method's power, sometimes dramatically, resulting in more rejections. Each step of our procedure is similar to the Bonferroni correction proposed by McCloskey (2012) and Romano, Shaikh and Wolf (2012), but, by choosing ϵ to be very small, we can embed the step in a sequential procedure and find more rejections.

Section 2 uses this principle to improve Romano and Wolf's (2005a) *StepM* procedure and increase its power for families of one-sided hypotheses. The StepM, like White's (2000) *Bootstrap Reality Check* (BRC) and Hansen's (2005) test of *Superior Predictive Ability* (SPA), uses the bootstrap to approximate the joint distribution of the test statistics for each hypothesis, and so obtains higher power than methods that assume a worst-case dependence structure (Holm, 1979, for example) and more general validity than those that assume

a convenient dependence structure. Romano and Wolf (2005a) improve on White (2000) and Hansen (2005) by incorporating an iterative stepdown method as described above; White (2000) and Hansen (2005) propose single step procedures. Our refinement amounts to using a heavily asymmetric two-sided version of the StepM and removing hypotheses far from the boundary between the null and the alternative in either direction, but then, after the sequential procedure stops, only rejecting the hypotheses that violate the null. This refinement is similar to existing procedures—Hansen (2005) proposes discarding the hypotheses with corresponding t-statistics below $-\sqrt{2\log\log n}$ before using the BRC for the null hypotheses $\theta_s \leq 0$, a threshold motivated by the Law of the Iterated Logarithm, and Hsu, Hsu and Kuan (2010) propose the same procedure for the StepM—but our threshold accounts for the number of hypotheses, giving it better size control in finite samples. Simulations presented in Section 4 show that Hansen's (2005) and Hsu, Hsu and Kuan's (2010) test can overreject in practice.

Section 2 also shows how to apply this procedure to test composite null hypotheses with several inequality restrictions (Corollary 1), and how to apply that result to the partial identification problem considered by Imbens and Manski (2004) (Remark 7). More simulations presented in Section 4 show that our procedure has roughly equal power to Andrews and Barwick's (2012a) preferred statistic (their AQLR) and to McCloskey's (2012) and Romano, Shaikh and Wolf's (2012) procedures to detect at least one violation of the inequalities. As mentioned earlier, our procedure (and McCloskey's, 2012, and Romano et al.'s, 2012) have the advantage of also controlling FWE, so the individual rejections can be taken as evidence against the individual hypotheses—in contrast, Andrews and Barwick's (2012a) AQLR only tells the researcher that one or more of the inequalities does not hold, but not which one. Our method has the further advantage that it will typically reject more of the individual false hypotheses than McCloskey's (2012) and Romano, Shaikh and Wolf's (2012), even though the probabilities of rejecting the composite null hypothesis are roughly equal.

Section 3 applies the same concepts to procedures that control other generalized error rates, namely the k-FWE and the False Discovery Proportion (FDP). These error rates can be used when FWE is too demanding a measure to be useful. In such situations, the researcher may be willing to allow for a few false rejections (k-FWE), or allow for a known percentage of the total rejections to be false (FDP, but see Section 3 for formal definitions of these terms). We show that the same ideas apply and present refinements to Romano and Wolf's (2007) k-StepM, a sequential procedure designed to control these error rates. In addition to corrections for one-sided testing, we

also present new restrictions that are implied by the error rates themselves and apply to families of two-sided tests as well.

Sections 2 and 3 lay out our theory as described above. Section 4 presents Monte Carlo simulations studying the behavior of our procedure and several competing methods in finite samples. Section 5 concludes.

2. Testing families of one-sided hypotheses with FWE control. Consider the following environment. Suppose that there are S null hypotheses $H_s : \theta_s \leq 0$ against the alternatives $H_s' : \theta_s > 0$, and let $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_S)$ be an estimator of θ with known distribution under the null. Let $I = \{s \in \{1, \ldots, S\} : \theta_s \leq 0\}$ index the true null hypotheses. A critical value q that controls FWE at level α satisfies

$$\Pr[\hat{\theta}_s \geq q \text{ for at least one } s \in I] \leq \alpha \tag{2}$$

while rejecting as many hypotheses in $I^c = \{1, \ldots, S\} \setminus I$ as possible (we will deal with studentized statistics in the actual results, but (2) is presented with unstudentized statistics for simplicity). This is a more stringent criterion than controlling the probability that (2) holds only when $I = \{1, \ldots, S\}$, which would be the focus if this were a test of the composite null hypothesis.

We will derive our results under the following high-level assumption.

ASSUMPTION 1. *Suppose that $\sqrt{n}\,(\hat{\theta} - \theta) \to^d N(0, V)$ where V is positive semi-definite; let $\hat{F}_n$ be a consistent estimator of the limiting distribution F of $(\sqrt{n}\,(\hat{\theta}_1 - \theta_1)/\hat{v}_1, \ldots, \sqrt{n}\,(\hat{\theta}_S - \theta_S)/\hat{v}_S)'$, where v_s is the (s, s) element of V, $\hat{v}_s$ is a consistent estimator of v_s, and each v_s is uniformly positive; and let $\hat{\psi}^* \sim \hat{F}_n$.*

Consistency of $\hat{F}_n$ in this context means that

$$\sup_{x \in \mathbb{R}^S} |\hat{F}_n(x) - F(x)| \to 0$$

in probability. Typically the distribution $\hat{F}$ can be estimated through a bootstrap. We assume asymptotic normality to simplify the presentation and the proofs, but it is not essential. Moreover, we work with studentized statistics to improve the procedure's performance (see Hansen, 2005, and Romano and Wolf, 2005a,b, 2010, among many others) but that assumption can be relaxed.

Theorem 1 presents our variation of the StepM for testing $\theta_s \leq 0$. Remarks follow the statement of the result and the proof is in the appendix.

THEOREM 1 (FWE control for one-sided hypotheses). *Suppose Assumption 1 holds,* $M_0 = \{1, \ldots, S\}$ *and* $\alpha \in (0, 1)$. *For each* $j = 1, 2, \ldots$ *do the following:*

1. *Set* p_j *to be the* ϵ *quantile of the distribution of* $\min_{s \in M_{j-1}} \hat{\psi}_s^*$, *with* $\epsilon > 0$.
2. *Set* q_j *to be the* $1 - \alpha$ *quantile of the distribution of* $\max_{s \in M_{j-1}} \hat{\psi}_s^*$.
3. *Set* $M_j = \{s : \sqrt{n}\,\hat{\theta}_s/\hat{v}_s \in [p_j, q_j]\}$ *and stop when* $M_j = M_{j-1}$ *or* $M_j = \emptyset$.

Let q *denote the last* q_j. *Then*

$$\limsup_{\substack{n\to\infty \\ \epsilon\to 0}} \sup_{\theta \in \mathbb{R}^S} \Pr_\theta[\sqrt{n}\,\hat{\theta}_s/\hat{v}_s > q \text{ for at least one } s \text{ such that } \theta_s \leq 0] \leq \alpha. \tag{3}$$

REMARK 1. To implement this method, we must set ϵ. If the quantiles are estimated with a bootstrap, ϵ can be set arbitrarily small by using the minimum of the bootstrap replications for p_j:

$$p_j = \min_{b=1,\ldots,B} \min_{s \in M_{j-1}} \hat{\psi}_{bs}^*$$

where $\hat{\psi}_{bs}^*$ is the sth element of the vector $\hat{\psi}_b^*$ and $\hat{\psi}_1^*, \ldots, \hat{\psi}_B^*$ are the bootstrap replications. This construction also implies that $\epsilon \to 0$ and so $p_j \to -\infty$ in probability (slowly) as $n \to \infty$. This is the method used in Section 4.

REMARK 2. This procedure differs from Romano and Wolf's (2005a) StepM in the p_j term—if we set $p_j = -\infty$ they are the same. This term fills the same role as Hansen's (2005) and Hsu, Hsu and Kuan's (2010) threshold, and if we set $p_j = -\sqrt{2 \log\log n}$ our algorithm becomes Hsu, Hsu and Kuan's (2010). Even though Hsu, Hsu and Kuan's (2010) threshold explicitly depends on n and diverges to $-\infty$ as n grows, p_j will typically be substantially farther from zero than $-\sqrt{2 \log\log n}$ because it explicitly accounts for the number of hypotheses (and $\sqrt{\log\log n}$ grows very slowly). This has size implications that can cause Hansen's (2005) and Hsu, Hsu and Kuan's (2010) statistics to overreject, as shown in Section 4.

REMARK 3. Bootstrap methods will often generate a distribution $\hat{F}_n$ that satisfies the assumptions of Theorem 1 under standard conditions (see, for example, Politis, Romano and Wolf, 1999).

REMARK 4. Occasionally one or more of the hypotheses will be logically related, in the sense that one can not hold without the other or is no longer

interesting without the other. An example is the problem of testing the family of null hypotheses $\mu_s \in [\mathrm{E}\, Z_s, \mathrm{E}\, Z_s + c_s]$, $s = 1, \ldots, S/2$, each μ_s and c_s is a known constant but $\mathrm{E}\, Z_s$ is unknown and estimated with the sample average $(1/n)\sum_{i=1}^n Z_{is}$. We can then write $\theta_{2s-1} = \mathrm{E}\, Z_s - \mu_s$ and $\theta_{2s} = \mu_s - \mathrm{E}\, Z_s - c_s$ for each s to put the problem in Theorem 1's notation.

But, if $\theta_{2s-1} \leq 0$ is rejected, the null hypothesis $\theta_{2s} \leq 0$ is no longer interesting: we will conclude that $\mu_s \notin [\mathrm{E}\, Z_s, \mathrm{E}\, Z_s + c_s]$ regardless of the results of the second test. Similarly, if $\theta_{2s-1} \leq 0$ is rejected then $\theta_{2s} \leq 0$ is no longer interesting.

In this case, if either $\theta_{2s-1} \leq 0$ or $\theta_{2s} \leq 0$ is rejected in step j, both should removed from M_j before calculating the next set of critical values. Doing so will increase power without changing the control of FWE for the interval hypotheses of interest. Failure to remove both hypotheses will not hurt the FWE either, but it will reduce power unnecessarily. This approach was introduced by Shaffer (1986) for the Bonferroni correction; see also Goeman and Solari (2010).

Also note that both hypotheses can be removed in the previous example only when one is larger than q_j, not when it is less than p_j.

Although the focus of this paper is on testing many individual hypotheses, the algorithm in Theorem 1 also provides an attractive test statistic for joint tests of several inequality restrictions. Corollary 1 formalizes this application.

COROLLARY 1 (Testing composite one-sided hypotheses). *Suppose Assumption 1 holds and implement Theorem 1's algorithm to produce q. Then*

$$\limsup_{\substack{n\to\infty \\ \epsilon\to 0}} \sup_{\theta \leq 0} \Pr_\theta[\max_s \sqrt{n}\,\hat{\theta}_s/\hat{v}_s > q] \leq \alpha \tag{4}$$

where the inequality $\theta \leq 0$ holds element by element.

REMARK 5. Note that the algorithm should continue as long as it rejects inequalities because the corresponding statistic is lower than p_j; removing these inequalities lowers the next upper bound, q_{j+1}. Obviously, there is no need to continue the algorithm once an individual statistic is greater than q_j when testing the composite null. However, if the researcher wants to interpret the individual rejections as well, continuing to reject as many hypotheses as possible (while still controlling FWE) is probably appropriate. See the next remark as well.

REMARK 6. Our Monte Carlo section, Section 4, shows that our procedure has comparable power to tests designed specifically for the composite null hypothesis (as studied by, for example, Andrews and Barwick, 2012a). Although there are theoretical reasons to believe that those dedicated tests may have a power advantage in principle, there is another reason to prefer tests that control FWE, even if they suffer a slight power disadvantage: interpretation of the results. We can interpret the individual rejections to learn which inequalities are violated if the test controls FWE, but not if it only controls size for the composite null. If the test recommended by Andrews and Barwick (2012a) rejects we do not learn which inequalities fail, but if our test rejects, we do.

If the power difference were large, one might proceed sequentially: first test the composite null using a dedicated test, then if it rejects, try to identify which individual inequalities fail by applying our test. A similar argument to those used in this paper imply that such a procedure would remain valid (see Rosenbaum, 2008, and Goeman and Solari, 2010). Our simulations, as well as those presented by Andrews and Barwick (2012a) and Romano, Shaikh and Wolf (2012), indicate that there is not a large power difference, so applying both procedures is not necessary in general. But doing so may be useful if a researcher wants to tailor a first-stage test of the composite null to have high power against a particular, application-specific alternative.

REMARK 7. Theorem 1 and Corollary 1 apply to many settings where the parameter of interest is only partially identified. As an example, consider Imbens and Manski's (2004) missing data problem: (Y_i, W_i) is an i.i.d. sequence for $i = 1, \ldots, n$; W_i is Bernoulli; and Y_i is bounded between 0 and 1 a.s. and is observed only when $W_i = 1$. The parameter of interest is $\mathrm{E}\, Y_i$ which must satisfy

$$\begin{aligned} \mathrm{E}\, Y_i &\geq \mathrm{E}(Y_i \mid W_i = 1) \Pr[W_i = 1] \\ \mathrm{E}\, Y_i &\leq \mathrm{E}(Y_i \mid W_i = 1) \Pr[W_i = 1] + (1 - \Pr[W_i = 1]). \end{aligned} \tag{5}$$

All of the quantities in (5) can be estimated from the data; the lower bound comes from setting $\mathrm{E}(Y_i \mid W_i = 0) = 0$ and the upper bound from $\mathrm{E}(Y_i \mid W_i = 0) = 1$. Note that $\mathrm{E}\, Y_i$ can not be estimated consistently without further assumptions on the distribution of Y_i given $W_i = 0$, assumptions that may be unrealistic if individuals self-select into the data set, but researchers can still estimate valid confidence intervals and conduct hypothesis tests without such assumptions.

To use Corollary 1 to test $\mathrm{E}\,Y_i = \mu_0$ for some value μ_0, we can define

$$\begin{aligned} \theta_1 &\equiv \mathrm{E}(Y_i \mid W_i = 1)\Pr[W_i = 1] - \mu_0 \\ \theta_2 &\equiv \mu_0 - \mathrm{E}(Y_i \mid W_i = 1)\Pr[W_i = 1] - (1 - \Pr[W_i = 1]), \end{aligned} \tag{6}$$

so (5) becomes $(\theta_1, \theta_2) \leq (0, 0)$. Also define

$$\hat{\theta}_1 = (1/n)\sum_{i=1}^{n} Y_i\, 1\{W_i = 1\} - \mu_0$$

$$\hat{\theta}_2 = \mu_0 - (1/n)\sum_{i=1}^{n} Y_i\, 1\{W_i = 1\} - \Big(1 - (1/n)\sum_{i=1}^{n} 1\{W_i = 1\}\Big).$$

Assuming $\Pr[W_i = 1]$ is bounded away from zero (as do Imbens and Manski, 2004) and standard moment and dependence conditions, each $\sqrt{n}(\hat{\theta}_i - \theta_i)$ is asymptotically normal under the null and Corollary 1 applies, even if $\Pr[W_i = 1]$ is near 1.

Confidence intervals for $\mathrm{E}\,Y_i$ can be constructed by inverting these hypothesis tests as usual. Notice that, when the gap between $\hat{\theta}_1$ and $\hat{\theta}_2$ is large, the confidence interval will be based on the distribution of only the closest $\hat{\theta}_i$ since the other inequality will be rejected with very high probability in the first stage, but when the gap is small the interval will use the distributions of both estimators, matching the key features of Imbens and Manski (2004).

3. Tests that control generalized error rates for families of one-sided and two-sided hypotheses. In some applications, tests that control FWE lack sufficient power and it may be appropriate to control a weaker measure of the error rate. In this section, we show how to apply the principles of the previous section to stepdown methods that control two such measures, the k-FWE and the False Discovery Proportion. We first consider k-FWE, a straightforward extension of the FWE. A critical value that controls k-FWE at level α satisfies

$$\Pr[\hat{\theta}_s \geq q \text{ for at least } k \text{ values of } s \text{ such that } \theta_s \leq 0] \leq \alpha \tag{7}$$

for one-sided tests or

$$\Pr[|\hat{\theta}_s| \geq q \text{ for at least } k \text{ values of } s \text{ such that } \theta_s = 0] \leq \alpha \tag{8}$$

for two-sided tests (as before, (7) and (8) use unstudentized statistics for simplicity, but our results will use studentized statistics for improved performance).

Stepdown procedures that control k-FWE face some new difficulties. By design, they continue to run after rejecting true hypotheses, so each step after the first operates under the assumption that some true hypotheses have been rejected, but fewer than k (meaning that the previous steps did not violate (7) or (8)). In Romano and Wolf's (2007) k-StepM procedure, separate critical values are generated using every subset that contains $k-1$ of the rejected hypotheses, and then the most conservative (largest) of those critical values is used for that step of the test. Even if k is relatively small (5 or 6) taking these combinations can be computationally costly.

Our algorithm improves on the k-StepM by ignoring the statistics so large that they would occur with negligibly small probability under the null. It also partitions the alternative space, further restricting the combinations of $k-1$ elements that must be calculated, by estimating the distribution of the ith largest $\hat{\theta}_s$ under the null, for every $i = 1, \ldots, k$—if both $\hat{\theta}_1$ and $\hat{\theta}_2$ are larger than the upper bound for the second-largest test statistic, combinations that include both $s = 1$ and $s = 2$ can be ignored (see the statements of Theorems 2 and 3, in particular the definition of R_j, and Remark 8 for a more precise statement). Theorem 2 presents a result for two-sided hypotheses, and Theorem 3 for one-sided hypotheses. For both results, define the k-max operator to return the kth largest of its arguments and let $\#A$ denote the number of elements in a set A. Remarks follow the statement of the theorem.

THEOREM 2 (k-FWE control for two-sided hypotheses). *Suppose Assumption 1 holds and let* $M_0 = \{1, \ldots, S\}$, $R_0 = \{\emptyset\}$, *and* $\alpha \in (0, 1)$. *For* $j = 1, 2, \ldots$ *do the following:*

1. *Set* r_{ij} *to be the* $1 - \epsilon$ *quantile of the distribution of*

$$\max_{I \in R_{j-1}} \; \underset{s \in M_{j-1} \cup I}{i\text{-max}} \, |\hat{\psi}_s^*|,$$

   with $i = 1, \ldots, k-1$ *and* $\epsilon > 0$.
2. *Set* q_j *to be the* $1 - \alpha$ *quantile of the distribution of*

$$\max_{I \in R_{j-1}} \; \underset{s \in M_{j-1} \cup I}{k\text{-max}} \, |\hat{\psi}_s^*|.$$

3. *Set* $M_j = \{s : |\sqrt{n}\,\hat{\theta}_s / \hat{v}_s| \leq q_j\}$,

$$N_{ij} = \begin{cases} \{s : |\sqrt{n}\,\hat{\theta}_s / \hat{v}_s| \in (r_{i+1,j}, r_{1j}]\} & i = 1, \ldots, k-2 \\ \{s : |\sqrt{n}\,\hat{\theta}_s / \hat{v}_s| \in (q_j, r_{1j}]\} & i = k-1 \end{cases}$$

   and

$$R_j = \{I \subset N_{k-1,j} : \#(I \cap N_{ij}) \leq i \text{ for } i = 1, \ldots, k-1\}.$$

Stop when $M_j = \emptyset$ *or*

$$(M_j, N_{1j}, \ldots, N_{k-1,j}) = (M_{j-1}, N_{1,j-1}, \ldots, N_{k-1,j-1}).$$

Again, let q *denote the last* q_j*. Then*

$$\limsup_{\substack{n\to\infty \\ \epsilon\to 0}} \sup_{\theta\in\mathbb{R}^S} \Pr_\theta[|\sqrt{n}\,\hat{\theta}_s/\hat{v}_s| > q \text{ for at least } k \text{ values of } s \text{ such that } \theta_s = 0] \le \alpha. \tag{9}$$

REMARK 8. As in most sequential algorithms, it is sufficient to show that each step of the algorithm controls the error rate only when all of the previous steps have already done so. So, for $j > 1$, we can assume that fewer than k true null hypotheses have been rejected in the previous steps. In Romano and Wolf's (2007) original proof, the outer maximum corresponding to our step 2 is taken over all sets I of size $k-1$, where $I \subset \{s : |\sqrt{n}\,\hat{\theta}_s/\hat{v}_s| > q_{j-1}\}$ (in our notation). The set of these I is potentially much larger than our R_{j-1}, reducing power and lengthening computational time.

We can use a smaller set because R_{j-1} removes the combinations that only occur with negligible probability under the null. For example, r_{1j} is essentially an upper bound on $\max|\sqrt{n}\ \hat{\theta}_s/\hat{v}_s|$ under the null and r_{2j} is essentially an upper bound on the second largest $|\sqrt{n}\,\hat{\theta}_s/\hat{v}_s|$ under the null. So at most one true hypothesis can have its test statistic between $r_{1,j-1}$ and $r_{2,j-1}$ and we can ignore all combinations that include the indices of two or more statistics between $r_{1,j-1}$ and $r_{2,j-1}$. The justification for the other bounds is the same.

Typically, the most useful restriction will be that statistics greater than r_{i1} can be rejected and ignored—none of the sets in R_{j-1} contain the indices of those statistics.

REMARK 9. As in Remark 1, if $\hat{F}_n$ is estimated with a bootstrap, it is often straightforward to set r_{ij} using

$$r_{ij} = \begin{cases} \max_{b=1,\ldots,B} \max_{s\in M_0} i\text{-max}\,|\hat{\psi}^*_{bs}| & j = 1,\ i = 1,\ldots,k-1 \\ \max_{b=1,\ldots,B} \max_{I\in R_{j-1}} i\text{-max}_{s\in M_{j-1}\cup I}\,|\hat{\psi}^*_{bs}| & j > 1,\ i = 1,\ldots,k-1 \end{cases}$$

where $\hat{\psi}^*_{1s}, \ldots, \hat{\psi}^*_{Bs}$ are the bootstrap replications of the test statistic for the sth hypothesis.

Theorem 2 can of course be modified for one-sided tests by adding the threshold p_j used in Theorem 1 (i.e. excluding those statistics that are too far below zero). Theorem 3 presents this result.

THEOREM 3 (k-FWE control for one-sided hypotheses). *Suppose Assumption 1 holds and let* $M_0 = \{1, \ldots, S\}$, $R_0 = \{\emptyset\}$, *and* $\alpha \in (0, 1)$. *For* $j = 1, 2, \ldots$ *do the following:*

1. *Set* p_j *to be the* ϵ *quantile of the distribution of*

$$\min_{I \in R_{j-1}} \min_{s \in M_{j-1} \cup I} \hat{\psi}_s^*,$$

   with $\epsilon > 0$.
2. *Set* r_{ij} *to be the* $1 - \epsilon$ *quantile of the distribution of*

$$\max_{I \in R_{j-1}} i\text{-}\max_{s \in M_{j-1} \cup I} \hat{\psi}_s^*,$$

   where $i = 1, \ldots, k-1$.
3. *Set* q_j *to be the* $1 - \alpha$ *quantile of the distribution of*

$$\max_{I \in R_{j-1}} k\text{-}\max_{s \in M_{j-1} \cup I} \hat{\psi}_s^*.$$

4. *Set* $M_j = \{s : \sqrt{n}\,\hat{\theta}_s / \hat{v}_s \in [p_j, q_j]\}$,

$$N_{ij} = \begin{cases} \{s : \sqrt{n}\,\hat{\theta}_s / \hat{v}_s \in (r_{i+1,j}, r_{1j}]\} & i = 1, \ldots, k-2 \\ \{s : \sqrt{n}\,\hat{\theta}_s / \hat{v}_s \in (q_j, r_{1j}]\} & i = k-1 \end{cases}$$

   and

$$R_j = \{I \subset N_{k-1,j} : \#(I \cap N_{ij}) \leq i \text{ for } i = 1, \ldots, k-1\}.$$

Stop if $M_j = \emptyset$ *or*

$$(M_j, N_{1j}, \ldots, N_{k-1,j}) = (M_{j-1}, N_{1,j-1}, \ldots, N_{k-1,j-1}).$$

Again, let q *denote the last* q_j. *Then*

$$\limsup_{\substack{n \to \infty \\ \epsilon \to 0}} \sup_{\theta \in \mathbb{R}^S} \Pr_\theta[\sqrt{n}\,\hat{\theta}_s / \hat{v}_s > q \text{ for at least } k \text{ values of } s \text{ such that } \theta_s \leq 0] \leq \alpha. \tag{10}$$

REMARK 10. Note that the parameters estimated to be far below the binding inequality are removed and do not enter as elements of $I \subset R_j$ or M_j. Just as before, this modification increases the test's power.

REMARK 11. If $\hat{F}_n$ is estimated with a bootstrap, we can often use

$$p_j = \begin{cases} \min_{b=1,\dots,B} \min_{s \in M_0} \hat{\psi}_s^* & j = 1 \\ \min_{b=1,\dots,B} \min_{I \in R_{j-1}} \min_{s \in M_{j-1} \cup I} \hat{\psi}_s^* & j > 1 \end{cases}$$

and

$$r_{ij} = \begin{cases} \max_{b=1,\dots,B} i\text{-}\max_{s \in M_0} \hat{\psi}_{bs}^* & j = 1,\ i = 1, \dots, k-1 \\ \max_{b=1,\dots,B} i\text{-}\max_{I \in R_{j-1}} i\text{-}\max_{s \in M_{j-1} \cup I} \hat{\psi}_{bs}^* & j > 1,\ i = 1, \dots, k-1 \end{cases}$$

where, again, $\hat{\psi}_{1s}^*, \dots, \hat{\psi}_{Bs}^*$ are the bootstrap replications of the test statistic for the sth hypothesis (also see Remarks 1 and 9).

We now turn to the second generalized error rate, the *False Discovery Proportion* (FDP). A critical value q controls FDP at level α if it satisfies

$$\Pr\left[\frac{\#\{s : |\hat{\theta}_s| \geq q \text{ and } \theta_s = 0\}}{\#\{s : |\hat{\theta}_s| \geq q\} \vee 1} > \gamma\right] \leq \alpha \tag{11}$$

for two-sided tests or

$$\Pr\left[\frac{\#\{s : \hat{\theta}_s \geq q \text{ and } \theta_s \leq 0\}}{\#\{s : \hat{\theta}_s \geq q\} \vee 1} > \gamma\right] \leq \alpha \tag{12}$$

for one-sided tests, for γ determined by the researcher in advance; i.e. it controls the probability that a predetermined percentage of the rejections are incorrect. As shown by Lehmann and Romano (2005), procedures that control k-FWE can be used to build procedures that control FDP. Suppose that a test that controls k-FWE at level α rejects N hypotheses. If $N > k/\gamma$, then $\Pr[k/N > \gamma] \leq \alpha$ and FDP is controlled at level α as well. So one can proceed sequentially in k, starting with $k = 1$, then 2, etc., stopping when $N \leq k/\gamma$. Corollary 2 demonstrates how to extend the algorithms presented in Theorems 2 and 3.

COROLLARY 2 (FDP control). *Suppose Assumption 1 holds and take* $\alpha, \gamma \in (0, 1)$.

1. *(One-sided hypotheses): Apply the algorithm of Theorem 3 sequentially at level* α *with* $k = 1, 2, \dots$ *producing a sequence of critical values* q_k*, and stop at the first* k *with*

$$k/\gamma \geq \#\{s : \sqrt{n}\, \hat{\theta}_s / \hat{v}_s > q_k\}. \tag{13}$$

Let q denote the last q_k. Then

$$\limsup_{\substack{n\to\infty\\ \epsilon\to 0}} \sup_{\theta\in\mathbb{R}^n} \Pr_\theta \left[\frac{\#\{s : \sqrt{n}\,\hat{\theta}_s/\hat{v}_s > q \text{ and } \theta_s \le 0\}}{\#\{s : \sqrt{n}\,\hat{\theta}_s/\hat{v}_s > q\} \vee 1} > \gamma \right] \le \alpha \tag{14}$$

2. *(Two-sided hypotheses): Apply the algorithm of Theorem 2 sequentially at level α with $k = 1, 2, \ldots$ producing a sequence of critical values q_k, and stop at the first k with*

$$k/\gamma \ge \#\{s : |\sqrt{n}\,\hat{\theta}_s/\hat{v}_s| > q_k\}. \tag{15}$$

Let q denote the last q_k. Then

$$\limsup_{\substack{n\to\infty\\ \epsilon\to 0}} \sup_{\theta\in\mathbb{R}^n} \Pr_\theta \left[\frac{\#\{s : |\sqrt{n}\,\hat{\theta}_s/\hat{v}_s| > q \text{ and } \theta_s = 0\}}{\#\{s : |\sqrt{n}\,\hat{\theta}_s/\hat{v}_s| > q\} \vee 1} > \gamma \right] \le \alpha \tag{16}$$

REMARK 12. The computational improvements of our algorithm are especially important when controlling FDP since k grows. To further reduce computational costs, step $k+1$ can be started where step k left off: if $r'_1, \ldots, r'_{k-1}$ and p' denote the last values of $r_{1j}, \ldots, r_{k-1,j}$ and p_j at step k, we can set $r_{i1} = r'_i$ and $p_1 = p'$ for step $k+1$.

REMARK 13. Note that steps can sometimes be skipped: if

$$(k+m)/\gamma < \#\{s : |\sqrt{n}\hat{\theta}_s/\hat{v}_s| > q_k\} \tag{17}$$

then we can go immediately to $k+m+1$ instead of $k+1$.

REMARK 14. If the computational costs become overwhelming, the algorithm can be stopped early. It still controls FDP at the prespecified levels, but obviously sacrifices some power. Since the computational costs grow with the number of hypotheses rejected, this scenario will come into play when many hypotheses have already been rejected and the loss of power may be acceptable.

4. Monte Carlo evidence. For a sense of the finite sample performance of our tests we present simulations for several different DGPs based loosely on Romano and Wolf's (2005a) Monte Carlo design. We study the performance of three of our procedures: the StepM modification for one-sided tests derived in Theorem 1, the k-StepM modification for one-sided tests described in Theorem 3, and the test of composite one-sided nulls described in Corollary 1. All of these simulations were programmed in R (R Development Core Team,

DGP	S	Out-performance	Under-performance	Equal Performance
1	2			$\mu_1 = \mu_2 = 1$
2	40			$\mu_1 = \cdots = \mu_{40} = 1$
3	40	$\mu_1 = \cdots = \mu_6 = 1.4$		$\mu_7 = \cdots = \mu_{40} = 1$
4	40	$\mu_1 = \cdots = \mu_6 = 1.4$	$\mu_7 = \cdots = \mu_{40} = -1$	
5	40	$\mu_1 = \cdots = \mu_{20} = 1.4$		$\mu_{21} = \cdots = \mu_{40} = 1$
6	40	$\mu_1 = \cdots = \mu_{20} = 1.4$	$\mu_{21} = \cdots = \mu_{40} = -1$	
7	4			$\mu_1 = \cdots = \mu_4 = 1$
8	4	$\mu_1 = \mu_2 = 1.4$		$\mu_3 = \mu_4 = 1$
9	4	$\mu_1 = \mu_2 = 1.4$	$\mu_3 = \mu_4 = -1$	

TABLE 1
Parameters for Monte Carlo experiments.

2012) and use the MASS (Venables and Ripley, 2002), xtable (Dahl, 2012), dbframe (Calhoun, 2010), RSQLite (James, 2012), R.Matlab (Bengtsson, 2013), and Combinations (Temple Lang, 2010) packages.

The Monte Carlo design is fairly basic: $\theta_s = \mathrm{E}\, X_{s,t} - \mathrm{E}\, Y_t$ for $s = 1, \ldots, S$. For design 1, $s = 2$ and

$$\begin{pmatrix} X_{1t} \\ X_{2t} \\ Y_t \end{pmatrix} \sim i.i.d.N \left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix} \right). \tag{18}$$

This covariance structure ensures that $\bar{X}_{1,\cdot} - \bar{Y}$ and $\bar{X}_{2,\cdot} - \bar{Y}$ are perfectly negatively correlated and this design is used to study size distortions in these procedures. For designs 2–6, $S = 40$ and

$$(X_{1,t}, \ldots, X_{40,t}, Y_t) \sim N(\mu, \mathrm{diag}(1, 2, 1, 2, \ldots, 2, 1)). \tag{19}$$

And for designs 7–9, $S = 4$ and

$$(X_{1,t}, \ldots, X_{4,t}, Y_t) \sim N(\mu, I). \tag{20}$$

Designs 1–6 are used for the multiple-testing simulations and designs 7–9 are used for the composite null hypotheses (one of the comparison methods, Andrews and Barwick's, 2012a, is computationally infeasible for 40 inequalities, so we drop the number). The mean, μ, is determined by the DGP; $\mathrm{E}\, Y_t = 1$ for all of the simulations, so $\mathrm{E}\, X_{i,t}$ takes on different values. Table 1 presents these different possible values.

The first Monte Carlo compares procedures that control FWE. It studies the size and power of Romano and Wolf's (2005a) StepM, Hsu, Hsu and Kuan's (2010) Step-SPA, and the our refinement of the StepM. The Step-SPA is a variation of Romano and Wolf's (2005a) StepM that initially discards

the null hypotheses s for which $\hat{\theta}_s/\hat{\sigma}_s \leq -\sqrt{2\log\log n}$, a threshold suggested by Hansen's (2005). All results are based on 1000 simulations and critical values are estimated using the i.i.d. bootstrap with 999 bootstrap samples, and use DGP designs 1–6. The lower thresholds, p_j, are set as the minimum of the bootstrap replications as suggested in Remark 1. The tests have nominal FWE of 5%, but simulations at 10% and 1% show similar patterns.

The second Monte Carlo compares procedures that control k-FWE with $k = 3$; it uses Romano and Wolf's (2007) k-StepM and our refinement presented in Theorem 2. The simulations use DGP designs 2–6 (the same as the first Monte Carlo, except that design 1 has fewer than k total hypotheses and is dropped) and the results are again based on 1000 simulations using an i.i.d bootstrap with 999 bootstrap samples, and the thresholds are set as in Remark 11 The nominal k-FWE is 5%.

The third Monte Carlo studies tests of composite null hypotheses: Andrews and Barwick's (2012a) AQLR and McCloskey's (2012) and Romano, Shaikh and Wolf's (2012) Bonferroni-based procedures in addition to our method described in Corollary 1; Andrews and Barwick (2012a) conduct an extensive simulation study in which they demonstrate that the AQLR performs better than many other recent tests of the composite null; see their paper for further discussion and comparisons. The Bonferroni-based method is a two-step procedure: it first conducts a one-sided test of the null $\theta_s \geq 0$ for each s at level $\alpha/10$, then constructs the $\alpha \cdot 9/10$ one-sided critical value for the null $\theta_s \leq 0$ for all s not rejected in the first-stage test. The procedure rejects if any of the test statistics are greater than this critical value, which can be approximated through the bootstrap. These simulations use DGP designs 7–9 for computational feasibility; the results are based on 1000 simulations and our method and the Bonferroni procedure both use 999 bootstrap samples. The AQLR is implemented using Matlab code provided by Andrews and Barwick (2012a,b) with their recommended settings, which is called from R via R.Matlab (Bengtsson, 2013). The nominal size for these tests is 5% and the lower threshold is again set as in Remark 1.

Table 2 presents the results for the FWE experiment and strongly supports this paper's new approach. Both the StepM and our refinement control FWE reliably, but the Step-SPA overrejects when there is a small number of equal-performing models—in simulation 1 it overrejects by almost 5 percentage points for 50 observations and 2.3 percentage points for 100 observations (note that the Step-SPA is equivalent to Hansen's original SPA in this experiment since all of the null hypotheses are true). For DGPs with no under-performing models (DGPs 1–3, and 5) our new procedure performs essentially the same as the StepM in terms of FWE and power. When some models under-perform

		Familywise error rate (%)			Average # discoveries			
# Obs.	Type	Ours	StepM	SPA	Ours	StepM	SPA	# False
50	1	5.0	5.0	9.8				0
	2	5.1	5.1	5.1				0
	3	1.6	1.6	1.6	0.8	0.8	0.8	6
	4	0.0	0.0	0.0	2.0	0.7	2.0	6
	5	3.2	3.2	3.2	2.7	2.7	2.7	20
	6	0.0	0.0	0.0	4.3	2.8	4.4	20
100	1	4.7	4.7	7.3				0
	2	4.6	4.6	4.6				0
	3	1.7	1.7	1.7	2.2	2.2	2.2	6
	4	0.0	0.0	0.0	4.1	2.1	4.1	6
	5	4.7	4.7	4.7	7.5	7.5	7.6	20
	6	0.0	0.0	0.0	10.7	7.7	10.7	20

TABLE 2

Results of the first Monte Carlo experiment—control of FWE. *The columns under the heading "Familywise error rate (%)" present results for our refinement of the StepM (under "Ours"), Romano and Wolf's (2005a) original StepM ("StepM") and Hsu, Hsu and Kuan's (2010) Step-*SPA *("SPA"). The columns under the heading "Average # discoveries" follow the same naming convention. The column "# False" lists the number of false hypotheses for that* DGP *for convenience. These results are based on 1000 simulations for each* DGP *using 999 bootstrap replications and the nominal* FWE *is 5%.*

		k-familywise error rate (%)		Average # discoveries		
# Obs.	Type	Ours	k-StepM	Ours	k-StepM	# False
50	2	4.1	4.1			0
	3	0.3	0.3	2.2	2.2	6
	4	0.0	0.0	4.2	1.7	6
	5	3.6	3.6	6.4	6.4	20
	6	0.0	0.0	10.3	6.6	20
100	2	4.6	4.6			0
	3	0.2	0.1	3.9	3.9	6
	4	0.3	0.0	5.6	3.5	6
	5	4.4	4.3	12.3	12.3	20
	6	0.0	0.0	16.4	12.5	20

TABLE 3

Results of second Monte Carlo experiment—control of k-FWE *with* $k = 3$. *The columns under the heading "k-familywise error rate (%)" present results for our extension of the k-StepM ("Ours") and Romano and Wolf's (2005a) original StepM ("k-StepM"). The columns under the heading "Average # discoveries" follow the same naming convention. The column "# False" lists the number of false hypotheses for that* DGP *for convenience. These results are based on 1000 simulations for each* DGP *using 999 bootstrap replications and the nominal* k-FWE *is 5%.*

		Size (%)			Power (%)			
# Obs.	Type	Ours	Bon.	AQLR	Ours	Bon.	AQLR	# False
100	1	4.7	4	4.4				0
	8	4.7	4	4.5				0
	9				86.9	85.3	86.9	2
	10				91.6	90.8	91.7	2

TABLE 4

Results of third Monte Carlo experiment—control of size when testing composites of one-sided hypotheses. The columns under the heading "Size (%)" present results for our method in Corollary 1 ("Ours"), McCloskey's (2012) and Romano, Shaikh and Wolf's (2012) Bonferroni-based procedure ("Bon."), and Andrews and Barwick's (2012a) AQLR *("AQLR"). The columns under the heading "Power (%)" follow the same naming convention. The column "# False" lists the number of false hypotheses for that* DGP *for convenience. These results are based on 1000 simulations for each* DGP *using 999 bootstrap replications and the nominal size is 5%.*

(DGPs 4 and 6), the new method identifies more incorrect null hypotheses than the original. For example, in DGP 4 with 100 observations, Romano and Wolf's (2005a) test finds on average 2.1 false hypotheses while this paper's test finds 4.1 out of 6, a substantial improvement; for 50 observations, the StepM finds 0.7 false hypotheses on average and this paper's test finds 2.0. Our method and the Step-SPA have basically the same power, but our method avoids over-rejecting when the inequalities bind.

Table 3 presents results for the k-FWE experiment, which again favor our approach. Our method and the k-StepM control the k-FWE at essentially identical (and correct) rates and when none of the models underperform the methods have almost identical power. But when some models do underperform, our method correctly rejects substantially more hypotheses. For example, in DGP 6 with 100 observations, our method rejects 4 more statistics (16.4 vs. 12.5) with identical control of k-FWE. The relative performance in other DGPs is similar.

Finally, Table 4 presents results for the size experiment. Here all of the methods perform about the same. All have estimated size slightly less than nominal size, but without cause for concern. And all of the statistics have nearly identical power when there are false hypotheses. As mentioned in Remark 6, an advantage of our statistic (and the Bonferroni-based statistic) is that researchers are justified in interpreting the individual statistic-by-statistic test results, while the AQLR does not. These simulations indicate that the power loss from taking this approach may be very small, the numbers are virtually identical, which makes our statistic more attractive.

Taken collectively, these simulations show that our improvements lead

to substantially more powerful tests in the multiple testing scenario that they were designed for, and also perform roughly as well as specialized (and complicated) statistics like the AQLR for testing composite null hypotheses.

5. Conclusion. This paper proposes simple modifications of existing stepdown procedures that increase power and reduce computational costs. The underlying idea—find and exclude events that occur with arbitrarily small probability in sequential testing—has other potential applications as well. Our simulation evidence indicates that the increase in power can be substantial.

APPENDIX A: PROOF OF MAIN RESULTS

PROOF OF THEOREM 1. Let $\{\theta_n\}$ be any sequence of vectors in $\mathbb{R}^S$ and $\{\epsilon_n\}_n$ a sequence of positive numbers that converges to zero as $n \to \infty$, where ϵ_n is used in place of ϵ in the theorem's statement. Then there exists a subsequence $\{n(m)\}_m$ of $\{n\}$ such that the limit of

$$\Pr_{\theta_{n(m)}}[\sqrt{n(m)}\,\theta_s > q \text{ for at least one } s \text{ such that } \theta_{n(m),s} \leq 0] \tag{21}$$

exists as $m \to \infty$; call this limit β. There also exists a further subsequence $\{n(m(\ell)\}_\ell$ such that each element of $\{\sqrt{n(m(\ell))}\,\theta_{n(m(\ell))}\}$ either converges to a finite limit or diverges to $\pm\infty$. To reduce the notational clutter, we'll write $n(m(\ell))$ as n_ℓ, $\epsilon_{n(m(\ell))}$ as ϵ_ℓ, and $\Pr_{\theta_{n_\ell}}$ as $\Pr_\ell$ for the rest of the proof. It suffices to prove that $\beta \leq \alpha$ for any such subsequence.

Define two subsets of $\{1, \ldots, S\}$:

$$I_1 \equiv \{s : -\infty < \lim_{\ell\to\infty} \sqrt{n_\ell}\,\theta_{n_\ell,s} \leq 0\}$$
$$I_2 \equiv \{s : \lim_{\ell\to\infty} \sqrt{n_\ell}\,\theta_{n_\ell,s} = -\infty\}.$$

We can assume that $I_1 \cup I_2$ is nonempty (otherwise $\beta = 0$ for this subsequence and the result is trivial). Moreover,

$$\lim_{\ell\to\infty} \Pr_\ell[\max_{s\in I_1\cup I_2} \hat{\psi}_s > q] \leq \lim_{\ell\to\infty} \Pr_\ell[\max_{s\in I_1} \hat{\psi}_s > q] + \lim_{\ell\to\infty} \Pr_\ell[\max_{s\in I_2} \hat{\psi}_s > q], \tag{22}$$

so it suffices to prove that

$$\lim_{\ell\to\infty} \Pr_\ell[\max_{s\in I_1} \hat{\psi}_s > q] \leq \alpha \tag{23}$$

and

$$\lim_{\ell\to\infty} \Pr_\ell[\max_{s\in I_2} \hat{\psi}_s > q] = 0 \tag{24}$$

and we can assume for the rest of the proof that neither I_1 nor I_2 are empty.

Start with the obvious inequality

$$\Pr_\ell[\max_{s\in I_1} \hat{\psi}_s > q] \le \Pr_\ell[\max_{s\in I_1} \hat{\psi}_s > q \text{ or } \min_{s\in I_1} \hat{\psi}_s < p]; \tag{25}$$

it suffices to bound the lim sup of the larger quantity. Consider the event on the right side of (25) and let j be the first index where one of the inequalities is violated. Then $\min_{s\in I_1} \hat{\psi}_s < p_j$ or $\max_{s\in I_1} \hat{\psi}_s > q_j$ must hold a.s.

We know (by construction of j) that $I_1 \subset M_{j-1}$ almost surely, and so $p_j \le p'$, and $q_j \ge q'$ almost surely where p' and q' are the ϵ_ℓ quantile of the distribution of $\min_{s\in I_1} \hat{\psi}^*_s$ and the $1-\alpha$ quantile of the distribution of $\max_{s\in I_1} \hat{\psi}^*_s$ respectively. Consequently,

$$\begin{aligned} \Pr_\ell[\max_{s\in I_1} \hat{\psi}_s > q \text{ or } \min_{s\in I_1} \hat{\psi}_s < p] &\le \Pr_\ell[\max_{s\in I_1} \hat{\psi}_s > q' \text{ or } \min_{s\in I_1} \hat{\psi}_s < p'] \\ &\le \Pr_\ell[\max_{s\in I_1} \hat{\psi}_s > q'] + \Pr_\ell[\min_{s\in I_1} \hat{\psi}_s < p'] \end{aligned} \tag{26}$$

and

$$\Pr_\ell[\max_{s\in I_2} \hat{\psi}_s > q] \le \Pr_\ell[\max_{s\in I_2} \hat{\psi}_s > q']. \tag{27}$$

Finally, consistency of $\hat{F}_n$ for the limiting distribution of $\hat{\psi}$ ensures that

$$\lim_{\ell\to\infty} \Pr_\ell[\max_{s\in I_1} \hat{\psi}_s > q'] \le \alpha, \tag{28}$$

$$\Pr_\ell[\min_{s\in I_1} \hat{\psi}_s < p'] \to 0 \text{ as } \ell \to \infty \tag{29}$$

and

$$\Pr_\ell[\max_{s\in I_2} \hat{\psi}_s > q'] \to 0 \text{ as } \ell \to \infty \tag{30}$$

completing the proof. □

PROOF OF THEOREM 2. As in the proof of Theorem 1, let $\{\theta_n\}$ be any sequence of vectors in $\mathbb{R}^S$ and $\{\epsilon_n\}_n$ a sequence of positive numbers that

converges to zero as $n \to \infty$, where ϵ_n is used in place of ϵ in the theorem's statement, and let $\{n_\ell\}_\ell$ and $\{\epsilon_\ell\}_\ell$ be subsequences such that, as $\ell \to \infty$,

$$\Pr_{\theta_{n_\ell}}[|\sqrt{n_\ell}\,\hat{\theta}_s/\hat{v}_s| > q \text{ for at least } k \text{ values of } s \text{ such that } \theta_{\ell,s} = 0] \to \beta \tag{31}$$

and each element of $\{\sqrt{n_\ell}\,\theta_{n_\ell}\}$ either converges to a finite limit or diverges to $\pm\infty$. It suffices to prove that $\beta \leq \alpha$ for any such subsequence.

Define $\hat{\psi}_s = \sqrt{n_\ell}\,\hat{\theta}_{n_\ell,s}/\hat{v}_s$ and write $\Pr_{\theta_{n_\ell}}$ as $\Pr_\ell$ for the rest of the proof to further simplify notation. Define a subset of $\{1, \ldots, S\}$:

$$I_1 \equiv \{s : \lim_{\ell\to\infty} \sqrt{n_\ell}\,\theta_{n_\ell,s} = 0\}.$$

We can assume that I_1 has k or more elements and it suffices to prove that

$$\lim_{\ell\to\infty} \Pr_\ell[k\text{-}\max_{s\in I_1} |\hat{\psi}_s| > q] \leq \alpha. \tag{32}$$

Note that

$$\Pr_\ell[k\text{-}\max_{s\in I_1} |\hat{\psi}_s| > q] \leq \Pr_\ell[k\text{-}\max_{s\in I_1} |\hat{\psi}_s| > q \text{ or } i\text{-}\max_{s\in I_1} |\hat{\psi}_s| > r_{i+1} \text{ for at least one } i = 1, \ldots, k-1]; \tag{33}$$

where each r_i denotes the last r_{ij}, so it suffices to bound the lim sup of the larger quantity. Let j be the first index where one of the inequalities is violated. Then

$$i\text{-}\max_{s\in I_1} |\hat{\psi}_s| > \max_{I\in R_{j-1}} r_{iI} \tag{34}$$

for some $i \in \{1, \ldots, k-1\}$ or

$$k\text{-}\max_{s\in I_1} |\hat{\psi}_s| > \max_{I\in R_{j-1}} q_I \tag{35}$$

must hold a.s., with r_{iI} the $1-\epsilon_\ell$ quantile of the distribution of $i\text{-}\max_{s\in M_{j-1}\cup I} \hat{\psi}^*_s$ and q_I the $1-\alpha$ quantile of $k\text{-}\max_{s\in M_{j-1}\cup I} \hat{\psi}^*_s$. We know (by construction of j) that $I_1 \subset M_{j-1} \cup I$ almost surely for at least one $I \in R_{j-1}$, and so

$$\max_{I\in R_{j-1}} r_{iI} \geq r'_i, \tag{36}$$

and

$$\max_{I\in R_{j-1}} q_I \geq q' \tag{37}$$

almost surely where each r'_i is the $1-\epsilon_\ell$ quantile of the distribution of i-max$_{s\in I_1}|\hat{\psi}^*_s|$ and q' is the $1-\alpha$ quantile of the distribution of k-max$_{s\in I_1}|\hat{\psi}^*_s|$. Consequently,

$$\begin{aligned}(38)\quad &\Pr_\ell[k\text{-}\max_{s\in I_1}|\hat{\psi}_s| > q \text{ or } i\text{-}\max_{s\in I_1}|\hat{\psi}_s| > r_i \text{ for at least one } i] \\ &\qquad \le \Pr_\ell[k\text{-}\max_{s\in I_1}|\hat{\psi}_s| > q'] + \sum_{i=1}^{k-1}\Pr_\ell[i\text{-}\max_{s\in I_1}|\hat{\psi}_s| > r'_i]\end{aligned}$$

and consistency of $\hat{F}_n$ ensures that

$$(39)\qquad \lim_{\ell\to\infty}\Pr_\ell[k\text{-}\max_{s\in I_1}|\hat{\psi}_s| > q'] \le \alpha$$

and

$$(40)\qquad \sum_{i=1}^{k-1}\Pr_\ell[i\text{-}\max_{s\in I_1}|\hat{\psi}_s| > r'_i] \to 0 \text{ as } \ell\to\infty$$

completing the proof. □

PROOF OF THEOREM 3. This proof is a straightforward combination of the arguments for Theorems 1 and 2 and is omitted. □

REFERENCES

ANDREWS, D. W. (2012). Similar-on-the-boundary tests for moment inequalities exist, but have poor power Discussion Paper No. 1815R, Cowles Foundation.

ANDREWS, D. W. K. and BARWICK, P. J. (2012a). Inference for parameters defined by moment inequalities: A recommended moment selection procedure. *Econometrica* **80** 2805-2826.

ANDREWS, D. W. K. and BARWICK, P. J. (2012b). Supplement to 'Inference for parameters defined by moment inequalities: A recommended moment selection procedure'. *Econometrica Supplemental Material* **80**.

BENGTSSON, H. (2013). R.matlab: Read and write of MAT files together with R-to-MATLAB connectivity R package version 2.0.1.

CALHOUN, G. (2010). dbframe: An R to SQL interface R package version 0.3.3.

DAHL, D. B. (2012). xtable: Export tables to LaTeX or HTML R package version 1.7-0.

GOEMAN, J. J. and SOLARI, A. (2010). The sequential rejection principle of familywise error control. *The Annals of Statistics* **38** 3782-3810.

HANSEN, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economic Statistics* **23** 365-380.

HIRANO, K. and PORTER, J. R. (2012). Impossibility results for nondifferentiable functionals. *Econometrica* **80** 1769-1790.

HOLM, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **6** 65-70.

Hsu, P.-H., Hsu, Y.-C. and Kuan, C.-M. (2010). Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. *Journal of Empirical Finance* **17** 471-484.

Imbens, G. W. and Manski, C. F. (2004). Confidence Intervals for Partially Identified Parameters. *Econometrica* **72** 1845-1857.

James, D. A. (2012). RSQLite: SQLite interface for R R package version 0.11.2.

Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *Annals of Statistics* **33** 1138-1154.

McCloskey, A. (2012). Bonferroni-Based Size-Correction for Nonstandard Testing Problems Working Papers No. 2012-16, Brown University, Department of Economics.

Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling*. Springer.

Romano, J. P., Shaikh, A. M. and Wolf, M. (2012). A simple two-step method for testing moment inequalities with an application to inference in partially identified models. Working Paper.

Romano, J. P. and Wolf, M. (2005a). Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica* **73** 1237-1282.

Romano, J. P. and Wolf, M. (2005b). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association* **100** 94-108.

Romano, J. P. and Wolf, M. (2007). Control of Generalized Error Rates in Multiple Testing. *The Annals of Statistics* **35** 1378-1408.

Romano, J. P. and Wolf, M. (2010). Balanced control of generalized error rates. *The Annals of Statistics* **38** 598-633.

Rosenbaum, P. R. (2008). Testing hypotheses in order. *Biometrika* **95** 248-252.

Shaffer, J. P. (1986). Modified Sequentially Rejective Multiple Test Procedures. *Journal of the American Statistical Association* **81** 826-831.

R Development Core Team (2012). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Temple Lang, D. (2010). Combinations: Compute the combinations of choosing r items from n elements. R package version 0.2-0.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*, 4th ed. Springer, New York. R package version 7.3.22.

White, H. (2000). A reality check for data snooping. *Econometrica* **68** 1097-1126.

Department of Economics
Iowa State University
Ames, Iowa 50011
E-mail: gcalhoun@iastate.edu