Um mapa para a transparência e replicabilidade na ciência social empírica: o Protocolo TIER

Amanda Domingos - Federal University of Pernambuco
Ian Rebouças Batista- Federal University of Pernambuco

Resumo

Transparência e replicabilidade são o padrão ouro da pesquisa científica. Mas, alcançar esse padrão pode ser um caminho longo e trabalhoso. Nesse artigo, apresentamos uma maneira de aumentar a transparência de sua pesquisa através da documentação. Argumentamos que o Protocolo TIER é uma importante ferramenta de transparência e de auxílio na organização de materiais para replicação. Trata-se de uma metodologia de documentação que compila os principais materiais produzidos numa pesquisa empírica social. Além disso, destacamos a adequação do protocolo nas três dimensões do padrão de replicação (substantiva, pedagógica e transparente), uma vez que contribui para o avanço da ciência, introduz jovens graduandas à análise de dados de uma forma prática e permite a transparência quanto aos procedimentos empregados na pesquisa. Por fim, discutimos atividades relacionadas ao Projeto TIER que contribuem com os objetivos do protocolo. Especificamente, apresentamos um exercício de introdução à análise de dados a ser realizado no R, onde a documentação do protocolo é requerida e sua aplicação, portanto, ilustrada.

Palavras-chave: Transparência; Replicabilidade; Documentação; Protocolo TIER

Abstract

Transparency and replicability are the gold standard in empirical research. But reaching that standard can be a long and tortuous journey. In this article, we present a way to increase the transparency of research through documentation. We argue that the TIER Protocol is an essential tool for transparency and assistance in organizing replication materials. It is a documentation methodology that compiles the main materials produced in empirical social research. Also, we highlight the adequacy of the protocol in the three dimensions of the replication standard (substantive, pedagogical, and transparent) – since it contributes to the advancement of science, introduces young undergraduate students to data analysis practically, and allows transparency of research procedures. Finally, we discuss activities related to the TIER Project that contribute to the objectives of the protocol. Specifically, we present a data analysis introduction exercise for R, where the protocol documentation is required and its application, therefore, exemplified.

Keywords: Replicability; Documentation; TIER Protocol

1. Introdução

Baixar o banco de dados, tratar o banco, elaborar *scripts*, programar funções, executar comandos, analisar os resultados. Se essas são etapas corriqueiras de sua pesquisa acadêmica, existe um protocolo que irá te ajudar a organizar a sua pesquisa e a divulgar, de maneira transparente e replicável, seus resultados.

Princípios e práticas de transparência e replicabilidade na pesquisa científica têm mobilizado crescente atenção da comunidade acadêmica nas ciências sociais, ganhando destaque em publicações nacionais (PARANHOS, 2012; PARANHOS et al, 2014; FIGUEIREDO et al, 2019; AVELINO e DESPOSATO, 2018) e internacionais (CHRISTENSEN, 2016; MUNAFÓ et al, 2017; CHRISTENSEN e MIGUEL, 2018; STOCKEMER, KOEHLER E LENZ, 2018; DE LA GUARDIA e STUDY, 2019), sendo ainda incentivados através de iniciativas inter-universitárias como *Berkeley Initiative for Transparency in the Social Sciences* (BITSS) e *Teaching Integrity in Empirical Research* (TIER), que promovem cursos, *workshops* e eventos voltados à troca de experiências e ideias em prol da transparência.

Identificada como padrão-ouro (KING, 1995), a transparência quanto aos procedimentos adotados em uma pesquisa é aqui entendida como a disponibilização dos materiais utilizados e a indicação das etapas percorridas até se chegar aos resultados. Para além dos benefícios que isso traz ao adequado funcionamento da ciência, permitindo acesso público a suas escolhas metodológicas e possibilitando que seu trabalho seja replicado e avaliado mais rigorosamente, a transparência permite: (i) teste de hipóteses mais preciso, (ii) a colaboração com outros pesquisadores com os mesmos interesses e (iii) aumento na credibilidade da autora. Argumentamos que a pesquisadora transparente, que documenta seu trabalho à medida que é feito, tem maior facilidade na elaboração de sua pesquisa. Uma documentação maior traz benefícios em termos de reputação científica, uma vez que contribui para boas práticas da ciência e gera benefícios práticos, como o aumento de citações (FIGUEIREDO et al, 2019).

A partir disso, alguns instrumentos têm sido empreendidos para difundir e implementar os princípios e as práticas da transparência e replicabilidade nas Ciências Sociais, sendo as três práticas principais: planos de divulgação, pré-registro e abertura de dados e materiais (MIGUEL et al, 2014). O objeto desse estudo se enquadro no terceiro tipo. Nesse artigo, apresentamos e sugerimos a adoção do protocolo TIER, uma metodologia para documentação adequada dos arquivos utilizados no desenvolvimento da pesquisa (bancos de dados, *scripts, outputs*) e posterior disponibilização desses materiais para replicação. Acreditamos que a adoção de uma regra específica quanto à organização dos arquivos de pesquisa aumenta a transparência e favorece a replicabilidade do estudo.

O artigo se divide como segue. Na próxima seção, apresentamos o padrão-ouro da ciência em termos de transparência dos procedimentos e replicabilidade dos resultados, indicando porque a leitora deve almejar esses valores e práticas. Na sequência, apresentamos o protocolo TIER, destacando os seus aspectos fundamentais. Em seguida, enfatizamos os benefícios e incentivos a se adotar uma estratégia de documentação e disponibilização para tornar a pesquisa transparente. Por fim, elencamos algumas atividades desenvolvidas pelo Projeto TIER que corroboram e buscam difundir as boas práticas de transparência e replicabilidade nas ciências sociais e apresentamos um exercício introdutório ao protocolo. As considerações finais encerram este trabalho.

2. Por que um Protocolo para transparência?

Dentre as características de uma boa pesquisa científica estão a inovação conceitual, o rigor metodológico

e a riqueza do conteúdo substantivo (PRZEWORSKI e SALOMON, 1998). Nas Ciências Sociais, no entanto, outros dois elementos foram adicionados a essas características nas últimas décadas: a transparência e a replicabilidade, tidos como o padrão-ouro da pesquisa empírica (KING, 1995; PARANHOS et al, 2014; JANZ, 2016). Os benefícios de uma pesquisa transparente são diversos, tanto para a disciplina como um todo quanto para a pesquisadora, em particular. A adoção de práticas transparentes permite que a roda da ciência gire de forma mais rápida, ao permitir o teste de hipótese mais preciso, facilitar a colaboração com outros pesquisadores com os mesmos interesses, aumentar a credibilidade da pesquisa, além de retirar a costumeira atenção da significância estatística dos resultados e aumentar a relevância da pesquisa *per se* (DE LA GUARDIA e STUDY, 2019).

No que se refere a benefícios para autores e periódicos, Figueiredo et al (2019) apresentam sete razões para adotar procedimentos transparentes em pesquisas empíricas nas ciências sociais: (I) prevenir que as pesquisas contenham erros honestos ou, até mesmo, fraudes em suas análises empíricas; (II) tornar a escrita de artigos científicos mais fácil, uma vez que permite o teste de outras hipóteses com a mesma base de dados; (III) procedimentos transparentes permitem que pareceristas tenham acesso a dados e *scripts* e deem recomendações mais precisas sobre como melhorar o trabalho; (IV) ao garantir a sistematização de dados brutos e analisados, permite que o trabalho cientifico do autor continue; (V) ajudam a construir reputação científica, ao demonstrar a boa-fé por apresentar material de replicação; (VI) ajudam a aprender análise de dados, ao permitir que os alunos tenham contato com dados e procedimentos estatísticos utilizados na vida real e (VII) aumenta o impacto do trabalho acadêmico, uma vez que evidências empíricas demonstram que trabalhos que adotam procedimentos e práticas transparentes são duas vezes mais citados.

Por outro lado, os custos da não adoção de procedimentos de transparência e replicabilidade também são variados. Pesquisas que não mantém procedimentos públicos podem contribuir para a diminuição da confiança em seus resultados, e, além disso, podem ser uma barreira à publicação em periódicos de alto fator de impacto que solicitam matérias de replicação – o que diminui o alcance da pesquisa (MUNAFO et al, 2017; DE LA GUARDIA e STUDY, 2019). Diante de tudo isso, uma pessoa poderia argumentar que não adota princípios e práticas de transparência por ser muito trabalhoso. Aqui, vale a pena reforçar que o padrão-ouro de replicação não solicita que um indivíduo reproduza os resultados apresentados em livros ou artigos científicos, apenas recomenda que sejam disponibilizadas tanta informação quanto for suficiente para que, caso algum pesquisador deseje, consiga reproduzi-los (KING, 1995).

Apesar disso, a prática de transparência não são a regra. Cerca de 67% dos artigos publicados na *American Political Science Review* entre 2013 e 2014 não são replicáveis (KEY, 2016). No Brasil, apenas 5% dos artigos publicados entre 2012 e 2016 em periódicos da área de Ciências Sociais eram passíveis de replicação completa (AVELINO e DESPOSATO, 2018)[1,2]. O padrão ainda é a obscuridade, não a transparência. Nesse sentido, não é incomum encontrar pesquisadores que não tornam públicos os procedimentos e práticas que os levaram às conclusões de seus produtos científicos. Questionamos a cara leitora: quantas vezes, na produção de sua própria pesquisa empírica, se esforçou para lembrar os caminhos que a levaram ao resultado que encontrou? Mais do que isso, o que você tem feito para ser transparente? Se pensou um tanto antes de responder essa última pergunta, esse artigo tem uma contribuição a te fazer. Nas próximas seções, vamos sugerir um protocolo

1 O relatório é fruto de do projeto "*Research Transparency in Brazilian Political and Social Science: A First Look*" dos autores junto à *Berkeley Initiative for Transparency in the Social Sciences* (BITSS). Para maiores informações, visite: < https://www.bitss.org/projects/research-transparency-in-brazilian-political-and-social-science-a-first-look/>. Acesso em: 20 mar. 2010.

2 O relatório final está publicamente disponível em: <https://osf.io/f279z/>. Acesso em 20 mar. 2020.

de organização para facilitar a implementação de práticas transparentes e discutir um pouco sobre o processo de documentação de uma pesquisa empírica.

3. O protocolo TIER

O protocolo TIER é uma metodologia de documentação dos materiais de replicação de uma pesquisa empírica. De modo simplificado, o protocolo traz uma indicação geral sobre a organização dos arquivos utilizados em um estudo, indica a construção de pastas e quais informações devem ser disponibilizadas em cada uma delas. O protocolo foi desenvolvido originalmente para lidar com um problema de sala de aula. Os trabalhos que eram entregues pelos alunos de Introdução à Estatística, disciplina ministrada por Richard Ball na graduação de economia na Haverford College (EUA), apresentavam problema similar àqueles existentes nos artigos científicos publicados nos periódicos acadêmicos de grandes revistas: eles não eram transparentes. Ao entregarem os trabalhos finais da disciplina com resultados errados, era praticamente impossível para o professor entender em que momento da atividade de análise dos dados foram realizados procedimentos equivocados, o que impedia a adequada correção do professor e, consequentemente, o processo de aprendizado do aluno.

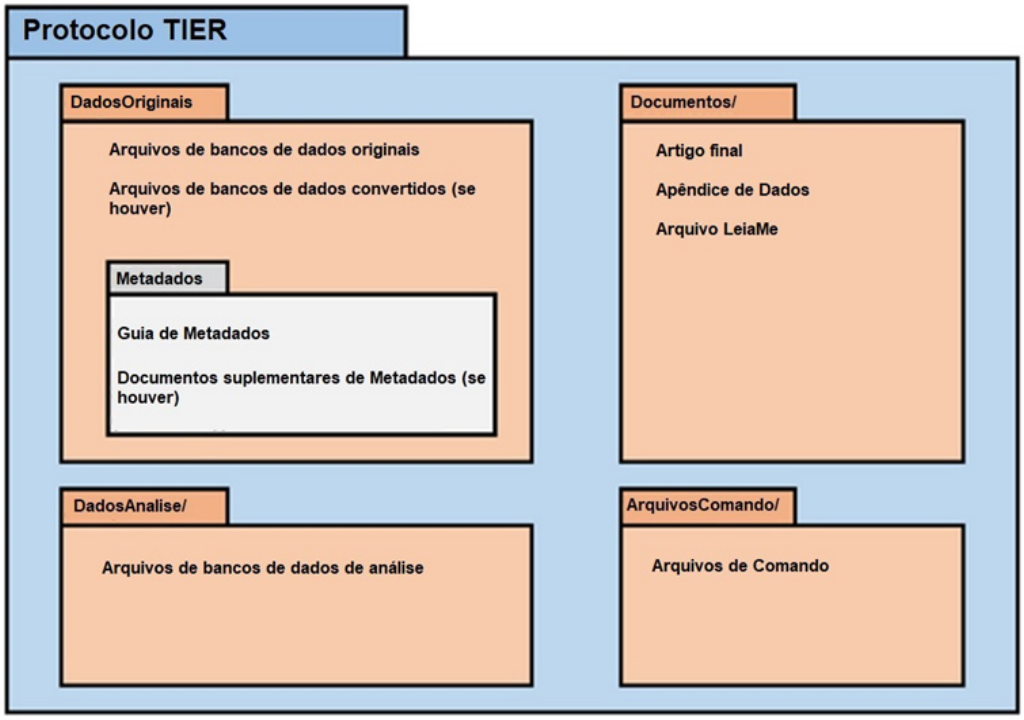
Para garantir que existisse um padrão para a entrega das atividades, tornando possível que o professor corrigisse de forma eficiente, tendo acesso aos *scripts* e dados de forma inteligível e organizada, Ball elaborou, junto com o bibliotecário do College, Norm Medeiros, o protocolo TIER (BALL e MEDEIROS, 2012). O protocolo consiste em uma organização e hierarquia de pastas específica que separa as informações sobre determinada pesquisa (dados brutos, dados analisados, *scripts*, etc.) e facilita a replicação dos achados. Apesar dessa origem pedagógica, o Protocolo TIER vai além das portas da sala de aula. Diversos pesquisadores na área de Ciências Sociais podem utilizar o protocolo como uma maneira de aumentar o nível de transparência de suas pesquisas. Nesse sentido, argumentamos que a adoção do Protocolo TIER acaba por colaborar nas três dimensões do padrão de replicação: substantiva, pedagógica e transparência (PARANHOS et al, 2013).

O protocolo TIER possui uma orientação geral, tal como se encontra ilustrado na Figura 1, e possui versões que aprimoram a organização e a operacionalização da replicação de seu trabalho, como a versão 4.0 apresentada na Figura 2. Discutiremos as duas versões na sequência, mas antes alguns esclarecimentos procedimentais dessa metodologia organizacional que o TIER propõe ajudam a entender sua proposta.

As diferentes pastas que irão compor o material de replicação de seu trabalho devem ser criadas no início da pesquisa. Elas acompanharão a pesquisadora ao longo de todo o processo: tratamento do banco de dados original, processamento das análises, elaboração do artigo e dos apêndices. A pesquisadora deve ir salvando e trabalhando com os arquivos nas pastas adequadas e referenciando, nos seus *scripts* ou arquivos de comando, os diretórios de trabalho de acordo com a hierarquia de pastas, o que torna fundamental que essa organização inicial seja mantida ao longo do trabalho. Por exemplo, no *script* onde você irá tratar o banco de dados original e irá gerar o banco da análise, você deve carregar o banco de dados original a partir da pasta *DadosOriginais* e deve salvar o banco tratado na pasta *DadosAnalise* (Figura 1). A ideia é que usando um *software* como o R, seu *script* contenha as linhas necessárias para abrir o banco original, que estará na pasta adequada, e salvar o banco de análise na pasta correta[3].

3 Como uma medida de replicabilidade, o ideal é que você adote um diretório de trabalho relativo. Para maiores informações, ver: https://ssc.wisc.edu/sscc/pubs/R_intro/book/1-10-paths-and-working-directories.html. Acesso em 07 de outubro de 2020.

Figura 1: Visão geral da documentação sugerida pelo Protocolo TIER



Fonte: Project TIER. Tradução dos autores4

Mas que organização de pastas você deve adotar, uma vez que as pastas propostas na Figura 1 e na Figura 2 são distintas? O objetivo do protocolo é fornecer uma orientação para organização e disponibilização de todo o material digital utilizado na elaboração de uma pesquisa empírica. É importante que essa orientação seja coerente e que seja, se não intuitiva, justificada e apresentada. Sendo assim, a exata configuração das pastas e dos arquivos fica a gosto do freguês. A Figura 1 é o exemplo mais genérico, a partir do qual se elaboraram quatro outras versões, e a Figura 2 é a formatação sugerida mais recente pelo Projeto (4.0). Contudo, se o argumento que o protocolo faz, sobre organização da pesquisa e sobre transparência, convencer a pesquisadora/leitora, você é encorajada a organizar o seu material para disponibilização a partir de um protocolo que funcione para você, ou seja, na sua própria hierarquia de pastas que achar conveniente.

Para que os protocolos individuais funcionem a qualquer pesquisador interessado em replicar o seu trabalho e conferir os resultados, são necessárias algumas orientações básicas que compõe a mensagem do TIER. O primeiro deles é a elaboração de arquivos LeiaMe (geralmente em .txt, para facilitar a leitura independente do sistema operacional utilizado) que orientarão a atividade de replicação desde o início. Uma das premissas de uma replicação transparente é que arquivos desse tipo são os primeiros a serem lidos, pois além de apresentar e explicar a configuração de pastas adotada, que a replicadora estará se deparando ao ler, são esses documentos de texto que dizem que arquivos abrir primeiro e qual será a sequência do trabalho de replicação. É interessante, portanto, que ele apareça logo na primeira pasta, como consta na Figura 2. A elaboração de um arquivo LeiaMe claro e objetivo acaba demonstrando a preocupação da pesquisadora com a replicação futura de seus achados. Assim, deve fornecer um pano-

4 O *template* utilizado é o *template* original utilizado pelo Projeto TIER. Disponível em: https://www.projecttier.org/tier-protocol/specifications/#overview-of-the-documentation. Acesso em 23 de março de 2020

rama completo da maneira com que as pastas estão organizas e das etapas de replicação.

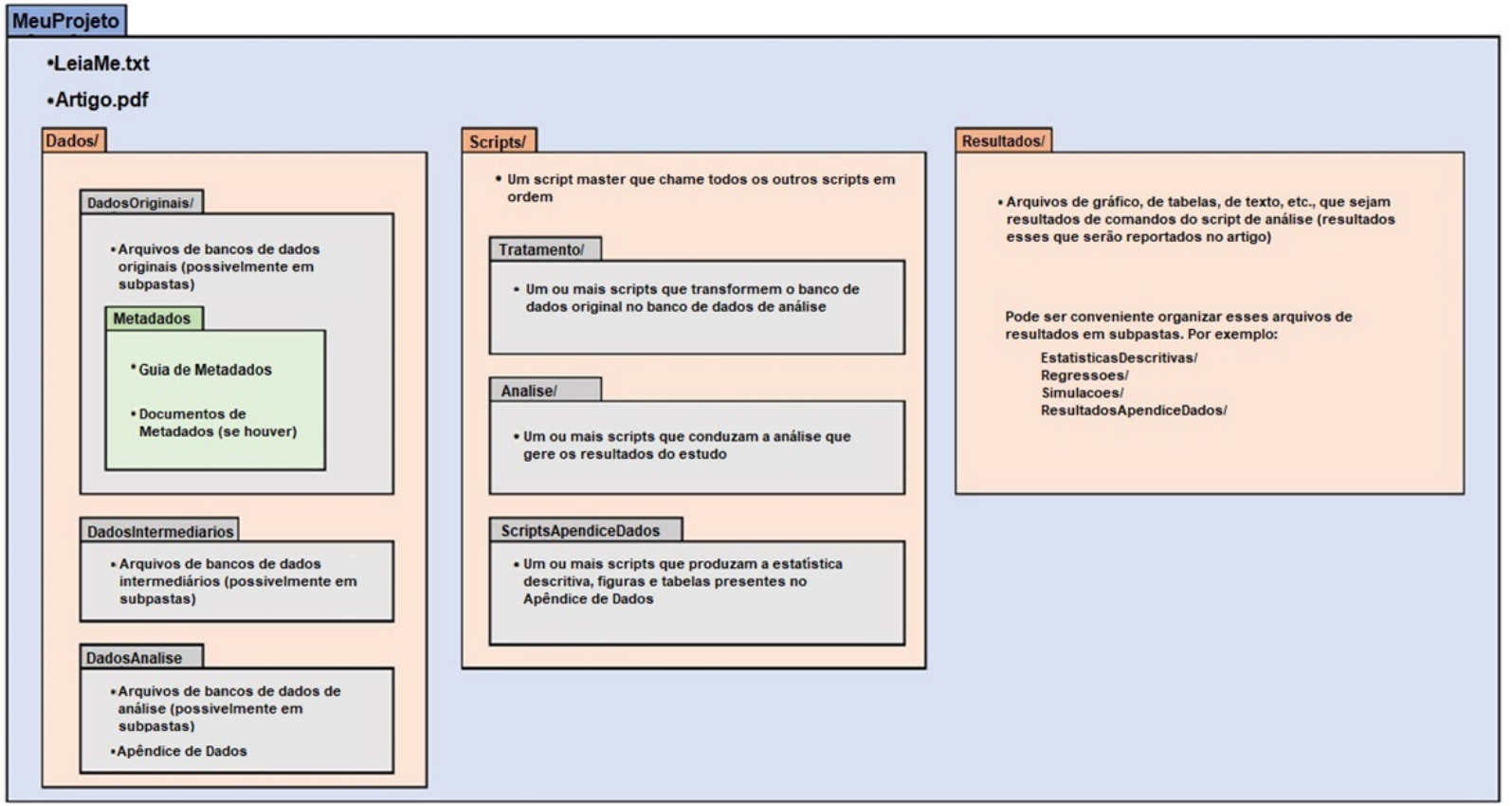
Por exemplo, para um trabalho que utiliza o protocolo inspirado no apresentado na Figura 2, o arquivo LeiaMe pode conter um texto semelhante a esse: "Inicialmente, vá até a subpasta *Tratamento*, localizada na pasta *Scripts,* e abra o *script* Tratamento.R. Nele você irá tratar o Banco.dta, que está na subpasta *DadosOriginais,* dentro da pasta *Dados*, e irá gerar o BancoNovo.dta, que será salvo na subpasta *DadosAnalise*. Em seguida, também dentro da pasta *Scripts*, vá até a subpasta *Analise*, abra o *script* Analise.R. Nele você irá gerar os Gráficos 1, 2 e 3, que serão salvos na pasta *Resultados*". Os arquivos LeiaMe podem ainda conter informações sobre o *software* de análise de dados utilizado, os endereços da *web* onde os bancos originais foram baixados ou ainda informações para contato com a pesquisadora.

Semelhante ao LeiaMe, mas com outra função, o protocolo TIER considera ainda que é fundamental que a pesquisadora elabore um Apêndice de Dados (ou *codebook*). É nesse documento que serão descritas com detalhes as variáveis que constam no banco de dados de análise, incluindo o tipo dessa variável, como ela é mensurada e como ela foi construída (se já estava pronta no banco de dados original ou se foi elaborada pela pesquisadora no processo de tratamento dos dados). Pode-se ainda incluir estatísticas descritivas para cada uma das variáveis. É nesse documento também que maiores especificidades sobre as versões do *software* utilizado e dos pacotes estatísticos necessários são fornecidas.

Outra orientação básica que permeia o protocolo TIER em qualquer de suas versões, e que deve estar presente também num possível protocolo pessoal inspirado no TIER, é a separação dos arquivos por tipos. Tanto na Figura 1 quanto na Figura 2 existem pastas próprias para bancos de dados, *scripts* e documentos. Analisemos de modo mais aprofundado essas figuras agora.

Na Figura 1, a orientação geral do protocolo, tem-se quatro pastas ao todo: *DadosOriginais*; *DadosAnalise*; *AquivosComando*; *Documentos*. Dentro da pasta *DadosOriginais*, encontramos os arquivos de bancos de dados originais, baixados diretamente da fonte que for, os bancos de dados convertidos, caso a pesquisadora tenha que converter o formato do banco de dados original baixado (note que mesmo nesse caso é importante manter o banco também no formato original), e uma subpasta chamada de *Metadados*, onde se encontram um Guia dos Metadados (um Apêndice de Dados para o banco original, incluindo a citação bibliográfica indicada, o identificador DOI, a data em que o banco foi baixado e o endereço de web) e documentos suplementares desse banco original (se houver). Na pasta *DadosAnálise*, tem-se o(s) banco(s) de dado(s) tratado(s), com o qual a pesquisadora fará as análises do seu artigo. Na pasta de *Documentos*, encontram-se o artigo final, o Apêndice de Dados e o arquivo LeiaMe. Por fim, na pasta *ArquivosComando*, estarão todos os *scripts* utilizados na pesquisa, desde o tratamento do banco de dados original até as análises realizadas.

Figura 2: Protocolo TIER na sua versão 4.0 (2020)



Fonte: Project TIER. Tradução dos autores

A Figura 2, por sua vez, baseia-se em uma elaboração preliminar da versão 4.0 do Protocolo, apresentada em novembro de 2019. Nela, na pasta principal, tem-se logo de cara dois arquivos, o artigo final e o arquivo LeiaMe, que irá orientar o procedimento daquele que deseja replicar o presente trabalho, e três pastas: *Dados*, *Scripts* e *Resultados*. Na pasta *Dados*, encontram-se três subpastas: *DadosOriginais* (com a pasta de *Metadados* dentro dela); *DadosIntermediarios*, onde estarão bancos de dados por ventura convertidos em outro formato; e *DadosAnalise*, onde também estará o Apêndice de Dados. Na pasta *Scripts*, encontra-se um *script* "*master*", que combinará todos os *scripts* do trabalho, mas também três subpastas: *Tratamento*, *Analise* e *ScriptApendiceDados*, cada pasta contendo o *script* que realiza cada uma das etapas referentes. Por fim, na pasta *Resultados*, estarão todos os arquivos que foram gerados ao longo da análise a partir dos *scripts*, como tabelas, gráficos e figuras, que podem, por sua vez, estar subdivididas também em subpastas.

Como se percebe, as duas versões do protocolo compartilham do mesmo ideal de documentação e transparência da pesquisa realizada, ainda que adotando organizações e hierarquias de pastas diferentes. Compreendendo quais os objetivos do protocolo e a lógica da separação dos arquivos presentes em ambas as versões, é possível então que a pesquisadora interessada numa pesquisa transparente e replicável possa, se não adotar exatamente alguma das versões expostas, adotar uma versão própria do protocolo TIER. Uma versão que faça sentido para a pesquisadora e beneficie o seu fluxo de trabalho, ao mesmo tempo que seja clara e acessível a qualquer interessado em replicar os resultados de sua pesquisa.

Após elaborado todo o material seguindo o Protocolo TIER, a disponibilização deve ocorrer da maneira mais acessível possível. Como a transparência no uso dos dados é um assunto em ascensão, temos a nosso dispor uma gama de repositórios digitais que fornecem a estrutura adequada para essa disponibilização. O *Open Science Framework*[5] (OSF), o *Dataverse*[6] e o *Github*[7] são três exemplos de repositórios adequados para o compartilhamento de todo o material de seu trabalho, inclusive com a possibilidade de divisão nas próprias pastas do protocolo. Nesse artigo, consideramos que o trabalho da pesquisadora tenha uma abordagem quantitativa, dada a natureza do protocolo que beneficia trabalhos deste tipo. Mas, também existem repositórios digitais para trabalhos com uma abordagem qualitativa[8], a saber: *Qualitative Data Repository*[9], UK data archive[10], National Anthropological Archives[11].

4. Documentação, protocolo TIER e dimensões do padrão de replicação

Em 1995, ano de lançamento do dossiê "*verification/replication*"[12], a realidade da produção científica nas ciências sociais era diferente. O dossiê propôs aos pesquisadores da área tentar tornar disponíveis maiores detalhes sobre as escolhas empíricas realizadas em suas pesquisas em notas de rodapé e apêndices metodológicos, mas esses esforços esbarravam nos limites impostos pelos periódicos e editoras de livro para divulgar tais detalhes. Por outro lado, a disponibilização de informações e armazenamento eram um problema: as facilida-

5 Ver: https://osf.io . Acesso em 20 mar. 2020.

6 Ver: https://dataverse.org/ . Acesso em 20 mar. 2020.

7 Ver: https://github.com/ . Acesso em 20 mar. 2020.

8 Para maiores informações sobre repositórios que aceitam dados qualitativos, ver: https://www.sesync.org/list-of-qualitative-data-repositories. Acesso em 24. Mar. 2020.

9 Ver: https://qdr.syr.edu/. Acesso em 24 mar. 2020.

10 Ver: https://www.ukdataservice.ac.uk/manage-data. Acesso em 24 mar. 2020.

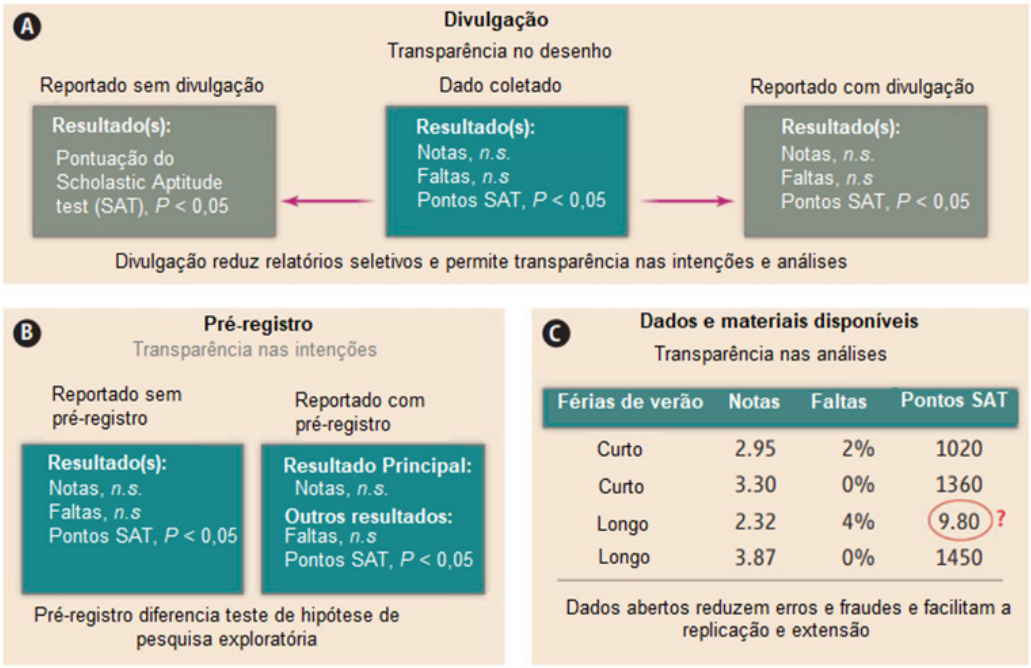
11 Ver: http://anthropology.si.edu/naa/search.html. Acesso em 24. mar. 2020.

12 O dossiê está disponível em: < https://www.cambridge.org/core/journals/ps-political-science-and-politics/issue/DAEAAA8412CD56791EAFCC75082A4C15>. Acesso em 24 mar. 2020.

des de *drives* e nuvens não existiam; o arquivo digital relacionado a publicações do *Inter-University Consortium for Political and Social Research* (ICPSR)[13], por exemplo, apenas seria lançado algum tempo depois do dossiê (ALVAREZ et al, 2018).

Nos dias atuais, com o surgimento de plataformas e repositórios *online* adequados, a disponibilização de informações tem se tornado cada dia mais fácil. Nesse sentido, diversos esforços têm sido empreendidos para difundir e implementar os princípios e as práticas da transparência e replicabilidade nas Ciências Sociais. Tais empenhos têm se concentrado em três práticas principais: planos de divulgação, pré-registro e abertura de dados e materiais (MIGUEL et al, 2014), conforme ilustrado na Figura 3.

Figura 3: Três mecanismos para aumentar a transparência nos relatórios científicos (Demonstração com a pergunta de pesquisa "as férias de verão mais curtas melhoram os resultados educacionais?", onde "n.s" significa p-valor > 0,05.)



Fonte: Miguel et al (2014), p.30. Tradução dos autores.

Os planos de divulgação surgem para advogar pela produção de relatórios sistemáticos que buscam garantir que os pesquisadores documentem e divulguem a forma com que procederam quanto a coleta e análise dos dados – essa é uma estratégia muito utilizada na área da medicina e implementada em alguns periódicos acadêmicos na Ciência Política e Psicologia[14] (MIGUEL et al, 2014). Por sua vez, o pré-registro é um documento público onde o pesquisador especifica variáveis, procedimentos de processamentos de dados e técnicas de análise (OLKEN, 2015). A adoção do pré-registro visa diminuir o viés de publicação[15], prevenir a apresentação de resultados seletivos da pesquisa e reforçar a credibilidade desta (CASEY et al, 2012). Por fim, a disponibilização de dados e materiais permitem que outros pesquisadores reproduzam os resultados da pesquisa, testem hipóteses remanescentes com os dados e identifiquem resultados não declarados ou, até mesmo, fraudes (MIGUEL et al, 2014; MUFANO et al, 2017; CHRISTENSEN e MIGUEL, 2018).

13 Disponível em: <www.icpsr.umich.edu/icpsrweb/deposit/pra/index.jsp>. Acesso em 24 mar. 2020.

14 De acordo com o levantamento dos autores, as revistas que adotaram políticas em direção a esse tipo de procedimento são: *Journal of Experimental Political Science*, *Management Science* e *Psychological Science*. Ver: Miguel et al (2014), p. 30.

15

É na terceira prática que o Protocolo TIER se enquadra. Mais especificamente, na documentação que precede a publicitação de dados e materiais de replicação. A adoção de políticas de replicabilidade por periódicos de alto fator de impacto forçam que as pesquisadoras documentem, de forma sistematizada, os arquivos de suas pesquisas. Apesar da demanda, a qualidade dos materiais disponibilizados não é alta. Stockemer, Koehler e Lenz (2018) analisaram todos os artigos publicados em 2015 nos três principais periódicos da área de comportamento político[16] e identificaram que em 25% deles os dados e os códigos estavam tão mal organizados que tornou a replicação impossível.

Materiais de replicação bem organizados aprimoram a extensão de pesquisas relacionadas e possibilitam que as análises sejam mais acessíveis (ALVAREZ et al, 2018).

No nível mais básico de preparação para se tornar um pesquisador transparente, é recomendado que se adote algum tipo de protocolo de organização desses dados, com o objetivo último de ter dados que sejam passíveis de reutilização e humanamente compreensíveis (FIGUEIREDO et al, 2019; OSINSKI, 2019).

Alvarez, Key e Nuñez (2018) apresentam algumas recomendações para aumentar a organização, clareza e utilização dos dados. Nele, os autores dão dicas sobre como arquivar os materiais relacionados à pesquisa de forma útil e segura, como produzir arquivos LeiaMe que sirvam como ajuda para replicações futuras, indicações sobre como devem ser nomeados os arquivos nos materiais de replicação, além de indicar como devem ser reportados *softwares* (utilizados ou criados pelo autor), *scripts* e *outputs* gerados na pesquisa. O Quadro 1 sumariza as sugestões dos autores.

16 Os autores analisaram todos os artigos presentes nos volumes de 2015 dos seguintes periódicos acadêmicos: *Electoral Studies*, *Party Politics* e *Journal of Elections, Public Opinion & Parties*.

Quadro 1 - Recomendações para criação de materiais de replicação (Alvarez et al, 2018)

Processos	Recomendações
Arquivamento	Os materiais de replicação devem ser armazenados em arquivos permanentes que estejam disponíveis por longos períodos. Atualmente, os mais confiáveis são os que têm garantias institucionais, sejam de consórcios ou universidades; Autores devem salvar seus arquivos de modo que possam ser utilizados no futuro (ex: arquivo separado por vírgula), ao invés de formatos software-specifc.
Arquivos LeiaMe	O arquivo deve ser claro e suficientemente detalhado, mas não em demasia. Alguns itens deveriam ser incluídos, a saber: (a) referência do artigo ou publicação associada; (b) uma breve explicação dos arquivos e tipos de arquivos incluídos, ex: dados originais, dados processados, scripts, etc; (c) uma indicação da ordem em que os scripts devem ser executados; (d) uma lista de softwares, pacotes de softwares e sistema operacional utilizado para produzir os resultados do artigo; (e) se extensões incomuns forem utilizadas, o autor deve indicar como proceder com esses arquivos.
Nomes dos arquivos	Os nomes dos arquivos devem ser fáceis de compreender e prover informação sobre seu conteúdo, isso é especialmente importante para matérias de replicação que incluem vários scripts e arquivos de dados. Caso existam arquivos que devem ser utilizados seguindo determinada ordem, eles devem ser nomeados de modo que incluía a ordem (ex: "1_DadosProcessados"; "2_Estimação", etc.).
Softwares e pacotes criados pelo usuário	Em casos onde o autor tenha criado um pacote estatístico, é necessário que estes estejam (idealmente) arquivados em um local estável, como o CRAN para o R. Se possível, o autor deve incluir uma cópia do software ou pacote nos materiais de replicação.
Compatibilidade com sistemas operacionais	Idealmente, é necessário que os autores garantam que seus materiais de replicação funcionem nos sistemas operacionais mais comuns. Caso não seja possível, devem indicar em que sistema eles funcionam.
Scripts	(a) devem conter uma breve discrição do que fazem, quais as dependências, quais pacotes são necessários e quais são os outputs; (b) deve-se evitar linhas sem especificação do que se faz, pois tal indicação ajuda a identificar partes especificas do gerenciamento e processamento de dados e a identificar erros ou possíveis problemas; (c) deve-se planejar bem e evitar recodificação de dados no meio da análise, ao menos quando seja extremamente necessário.
Resultados	Os materiais de replicação devem conter tabelas, gráficos e figuras da mesma forma que aparecem no artigo.
Resultados excluídos	Geralmente, materiais de replicação apresentam scripts ou dados que não foram apresentados no artigo ou que estavam vinculados a um apêndice online. Nesse caso, o autor deve diferenciá-los dos que estão no artigo.
Resultados intermediários	Algumas vezes os artigos apresentam resultados intermediários que levam a algo essencial na pesquisa. Nesse caso, devem estar presentes nos materiais de replicação.
Sequências aleatórias	Para estudos de simulação ou estratégias de estimativa que usem randomização, os autores devem sempre incluir a sequência aleatória de números usada para produzir os resultados no artigo.
Diretórios e pastas	Os caminhos de diretórios devem ser facilmente identificáveis no script, facilitando a alteração pelos usuários. É recomendado que se coloque o diretório apenas no início, modificá-los ao longo do script pode gerar confusão.
Computing time	Os autores devem indicar (no código e no arquivo LeiaMe) se algum script específico demora muito tempo para calcular. Isso é importante para que o usuário não se depare com uma estimativa longa e acredite que esteja passando por algum problema.
Parallel Computing	Cada vez mais os autores estão utilizando parallel computing nos seus scripts a fim de acelerar o processo computacional. Infelizmente, isso não funciona em todos os computadores. Logo, devem indicar em que momento do script a estratégia foi utilizada.
Avisos e erros	Não é incomum encontrar scripts que produzam erros ou avisos – que geralmente estão relacionados ao uso inadequado de pacotes ou compatibilidade com o sistema operacional. Nesse caso, os autores devem indicar as razões pelas quais eles não são um motivo de preocupação.

Fonte: Alvarez et al (2018), p. 425-426. Tradução dos autores

As sugestões de Alvarez et al. (2018) são, definitivamente, um caminho para a produção de uma pesquisa mais transparente e replicável. Apesar disso, a quantidade de passos necessários e a falta de um direcionamento mais específico sobre como implementá-los acaba tornando-as um caminho trabalhoso. É nesse sentido que defendemos a adoção do protocolo TIER como uma ferramenta para a organização de materiais de replicação.

Acreditamos que a adoção e ampla divulgação do Protocolo TIER oferece um benefício claro à comunidade acadêmica: ao oferecer uma maneira de padronizar os materiais de replicação, acaba fazendo com que pesquisadores e diferentes áreas saibam, de forma intuitiva, o que devem encontrar em cada uma das pastas. Em outras palavras, a organização dos materiais de replicação se tornará mais orgânica para os pesquisadores e mais intuitivas para os replicadores. Nesse sentido, o protocolo TIER surge como um caminho menos tortuoso à adoção do padrão de replicabilidade.

Tais benefícios estão amplamente fincados nas três dimensões da replicabilidade, a saber: substantiva, pedagógica e transparente. Na dimensão substantiva, o padrão de replicação possibilita que os trabalhos possam ser falseados e, com isso, que a ciência avance mais rápido (PARANHOS et al, 2012; PARANHOS et al, 2013). De modo similar, a adoção do Protocolo TIER permitiu que os trabalhos dos alunos de Ball fossem avaliados de forma mais cautelosa e que os alunos recebessem comentários mais direcionados sobre os problemas enfrentados na produção do artigo. Por sua vez, para os pesquisadores que resolvem adotar o protocolo, esta estratégia de organização permite que seus trabalhos recebam melhores avaliações pelos seus pares e pelos avaliadores de periódicos especializados.

No que diz respeito à dimensão pedagógica, o padrão de replicação permite que os alunos sejam iniciados no mundo da pesquisa e da análise de dados ao terem contato com dados e problemas do mundo real (KING, 1995). Mais recentemente, os alunos de Ball resolvem exercícios *soup-to-nuts* (mais sobre eles abaixo), que levam os alunos a implementar o protocolo e exercer todo o processo de pesquisa empírica com dados estatísticos, desde a coleta dos dados à escrita do relatório com as análises estatísticas. Nesse mesmo sentido, o *hub* do Projeto TIER no Brasil, baseado em Recife, tem iniciado a produção de *workshops* para alunos de graduação em Ciência Política e Relações Internacionais com o objetivo de difundir a utilização do protocolo e iniciá-los na pesquisa empírica.

Por fim, na dimensão da transparência, o padrão de replicação permite que os procedimentos e dados sejam publicamente disponíveis para todos os pesquisadores que se interessem sobre o tema (KING, 1995; PARANHOS et al, 2013; FIGUEIREDO, 2019). O protocolo permite que a disponibilização das informações necessárias para a replicação seja humanamente inteligível e disponibilizada de modo com que qualquer pesquisador possa realizar a replicação dos resultados do estudo sem necessitar de ajuda do autor. Para os alunos do Ball, possibilitou que os resultados fossem replicados sem a presença dos alunos. Para a ciência, é a maneira das pesquisas, discussões e hipóteses conflitantes avançarem rumo à produção do conhecimento.

Nesse sentido, acreditamos que a adoção do protocolo pode ser uma ferramenta útil na caminhada para os autores das Ciências Sociais que desejam alcançar maiores níveis de transparência em suas pesquisas, auxiliando não só quem está conhecendo o valor da transparência agora, mas também fornecendo um mapa claro sobre como proceder para aqueles que já tinham o interesse, mas não sabem como. Na próxima seção, apresentamos um exercício *soup-to-nuts* de replicação que serve como uma introdução para aplicação do protocolo.

5. Exercício de Aplicação do Protocolo

Nesta seção, apresentaremos o primeiro exercício *soup-to-nuts* desenvolvido para exemplificação da aplicabilidade do protocolo (BALL, 2018). Os exercícios *soup-to-nuts* são originalmente desenvolvidos com o intuito de serem aplicados em turmas de Introdução à Estatística ou Introdução à Análise de Dados[17]. Nessas atividades, o intuito é levar os alunos a realizarem um processo completo de pesquisa empírica, desde o download de um banco de dados, até o tratamento, análise e entrega dos resultados. A descrição das etapas é detalhada, e originalmente é prevista para ser introduzida em sala de aula e concluída pelos alunos em casa. Nessa atividade, os alunos devem construir a sua hierarquia de pastas logo no início do projeto, trabalhando, portanto, com o Protocolo TIER ao longo de toda a jornada, uma vez que serão cobrados de realizar a entrega dos resultados com a documentação completa e organizada. Além de facilitar o trabalho do professor na correção dos exercícios, como já comentamos da experiência do Richard Ball, o protocolo ajuda aos alunos em seus fluxos de trabalho, configurando assim um exemplo poderoso das potencialidades dessa metodologia organizacional. Os exercícios de *soup-to-nuts* são, portanto, verdadeiros tutoriais para os alunos trabalharem com análise de dados de maneira transparente e replicável.

Neste artigo, utilizaremos uma versão encurtada do exercício "Consumo de Álcool em Universidades Americanas" como exemplificação do protocolo[18]. Recomendamos que dedique algum tempo para sua elaboração, de maneira que compreenda a maneira pela qual o protocolo contribui ao fluxo de trabalho e facilita a replicação de uma pesquisa. A versão completa e em português deste exercício você encontra na página do OSF[19], e a versão original você encontra no site do TIER[20]. O exercício resolvido, com modelo de relatório e de *scripts*, também estão disponíveis nos endereços indicados. A versão apresentada abaixo é para ser resolvida no *software* R.

Atividade de Replicação (Protocolo TIER): Consumo de Álcool em Universidades Americanas

Nesta atividade, você irá explorar se o consumo de bebidas alcoólicas é menor entre estudantes que vivem em dormitórios onde a prática é proibida em relação àqueles que moram em dormitórios onde não existe essa regra. Os dados utilizados são fruto de uma pesquisa conduzida em 2001 em universidades americanas, da *Harvard School of Public Health College Alcohol Study, 2001*. Essa atividade pedirá que você use os dados dessa pesquisa para criar alguns gráficos de barra e que responda a uma série de questões sobre a interpretação desses gráficos no que se refere aos padrões de consumo de álcool entre estudantes, comparando os que podem beber em seus dormitórios e os que não deveriam. A entrega deverá ser não somente o relatório com os gráficos e as respostas, mas todo o material utilizado para processamento e análise realizados nesta atividade, seguindo a documentação indicada.

I. PREPARATIVOS

IA. Crie uma hierarquia de pastas para guardar e organizar seu trabalho.

Crie uma pasta chamada de **ExercicioAlcool/.**

17 Ver: https://www.projecttier.org/tier-classroom/soup-nuts-exercises/#shorter-exercises-for-teaching-transparency-and-reproducibility . Acesso em 20 de março de 2020.

18 Ball (2018) é o autor da versão original deste exercício, e Jenna Krall é creditada enquanto colaboradora. A tradução para o português e para a linguagem R foi realizada pelos membros do TIER hub Recife. Esta versão encurtada é de responsabilidade dos autores deste artigo, com autorização do autor da versão original.

19 Versão completa do exercício em Português: https://osf.io/mc2e9/?view_only=32e9fa640c83422ba7d6e9ab91e65743. Acesso em 06 de outubro de 2020.

20 Versão original completa: https://www.projecttier.org/tier-classroom/soup-nuts-exercises/#shorter-exercises-for-teaching-transparency-and-reproducibility . Acesso em 10 de outubro de 2020.

Dentro de **ExercicioAlcool/**, crie as seguintes pastas e subpastas:

- **Dados/** (uma subpasta de **ExercicioAlcool/**)
  - **DadosOriginais/** (uma subpasta de **Dados/**)
  - **DadosAnalise** (uma subpasta **Dados/**)
- ***Scripts/*** (uma subpasta de **ExercicioAlcool/**)
- ***Graficos/*** (uma subpasta de **ExercicioAlcool/**)

Salve essa hierarquia de pastas no local que escolheu guardar o trabalho para esse exercício.

I.B Instale os pacotes

Dois pacotes serão úteis nessa atividade. Instale antes de iniciar o seu trabalho.

- *haven* (para ler o banco de dados no formato *.dta*)
- *tidyverse* (para facilitar os procedimentos de processamento e criação dos gráficos)

II. BAIXE E EXAMINE O BANCO DE DADOS E O *CODEBOOK* QUE VOCÊ UTILIZARÁ NESSE EXERCÍCIO.

II.A. Encontre e explore o site onde os dados para esse exercício estão disponíveis.

Os dados da pesquisa está arquivado no *Inter-University Consortium for Political and Social Research* (ICPSR).

Vá até o site do ICPSR (www.icpsr.umich.edu) e procure por *Harvard School of Public Health College Alcohol Study, 2001*.

Quando você tiver encontrado a página desse estudo, leia as informações fornecidas.

II.B. Baixe o banco de dados e o *codebook*.

- Antes de baixar o banco de dados, você precisará criar uma conta no site do ICPSR.

Uma vez feito isso, faça o login e vá para a página do estudo: https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/4291

Clique no botão de *download*. Você verá um menu de formatos para escolher, incluindo SAS, SPSS, Stata ou ASCII. Nesse menu, selecione Stata (com o pacote *haven* o R irá ler o formato *.dta*).

O arquivo do banco será baixado numa pasta zipada, *ICPSR_04291-V2.zip*.

Quando você extrair os arquivos, verá que são vários documentos, incluindo uma subpasta chamada **DS0001**. Para esse exercício, você usará somente dois desses arquivos que vieram na pasta zipada e ambos estão na subpasta **DS0001**:

- *04291-0001-Data.dta*. Esse arquivo contém os dados da pesquisa em formato *.dta*. Esse arquivo será referido como "o banco de dados original" nesse exercício.
- *04291-0001-Codebook.pdf*. Esse é o *codebook* para a pesquisa utilizada. Será referido como "metadados do banco de dados original".

Você deve salvar cópias desses dois arquivos na pasta **DadosOriginais/**.

II. C. Examine o *codebook* e o banco de dados.

Examine o *codebook* e o banco de dados original para se familiarizar com as informações que eles contêm. Para esse exercício, você precisará utilizar cinco variáveis desse estudo. No banco de dados original, essas variáveis têm os nomes *A6, B8, B9, C13* e *G11*. Para cada uma dessas variáveis, analisando o *codebook*, responda as seguintes questões (as respostas a essas perguntas não precisarão estar no relatório final, mas são fundamentais para que você compreenda a natureza dos dados com que lidará):

- Como a variável é definida? O que ela representa?
- Qual a classe da variável (contínua, ordinal, categórica, discreta?)
- É qualitativa ou quantitativa?
    - Se for quantitativa, encontre a média, o desvio padrão, os três quartis, o valor mínimo e o valor máximo.
    - Se for qualitativa categórica, quantas categorias existem, o que elas representam e qual proporção das observações por categorias? Para cada variável categórica, decida se as categorias estão ordenadas ou não.

III. ESCREVENDO OS *SCRIPTS*

Convenções para o diretório de trabalho

Em todos os *scripts*, adote as seguintes convenções para o diretório de trabalho:

- Quando seus *scripts* forem executados, o diretório de trabalho do R deve ser definido para a pasta ***Scripts/***.
- Não escreva comandos em seus *scripts* que mudem o diretório de trabalho. Por exemplo, evite utilizar a função *setwd()*.
- Quando você precisar abrir ou salvar um arquivo em outra pasta que não ***Scripts/***, especifique a localização da pasta em questão utilizando um *relative directory path*.

Para esse exercício, você escreverá três *scripts*:

III.A. *tratamento.R*

- O propósito desse *script* é modificar o banco original (*04291-001-Data.dta*) para prepará-lo para a análise que você vai realizar.
    - As modificações incluem a criação de uma série de novas variáveis e a exclusão de algumas observações do banco.
- Salve o novo banco em um arquivo chamado *analise.RData*.
    - Esse arquivo será chamado de banco de análise.

Criando o *script*

- Abra o RStudio.
- Abra um novo *script* de R.
- No início do novo *script*, escreva um comentário indicando que o diretório de trabalho do R deve ser a pasta **Scripts/**. Por exemplo:

  # Quando esse *script* for executado, o diretório de trabalho deve ser a pasta **Scripts/**.
- Salve esse novo *script* na sua pasta **Scripts/**, com o nome *tratamento.R*.
    - Para evitar perder trabalho, salve o *scripts* constantemente enquanto o escreve.
- Carrege os pacotes *haven* e *tidyverse*
- Abra (importe) o banco *04291-0001-Data.dta*.
    - Use a função *read_dta()* do pacote *haven*.
    - Como seu diretório de trabalho é a pasta **Scripts/**, você precisa usar um *relative directory path* para especificar que o arquivo *04291-0001-Data.dta* deve ser importado de **DadosOriginais/**.
- Para essa atividade, nós queremos considerar apenas estudantes que vivem no campus ou em fraternidades/sororidades. Assim, exclua todos os casos os quais a residência do aluno seja "*Off*

campus house/apt" ou "*other*" ("fora de casa/apto no campus" ou "outro"). Também exclua todos os casos onde a variável representando a residência do aluno seja uma NA.

- Crie uma variável chamada *drunk*, que seja:
    - Igual a 0 se o estudante bebeu o suficiente para ficar bêbado menos do que três vezes nos últimos trinta dias.
    - Igual a 1 se o estudante bebeu o suficiente para ficar bêbado três ou mais vezes nos últimos trinta dias.
    - Igual a NA se a variável indicando quantas vezes o estudante bebeu o suficiente para ficar bêbado nos últimos trinta dias for NA.
- Crie uma variável chamada *hsdrunk*, que seja:
    - Igual a 0 se o estudante bebeu cinco ou mais bebidas seguidas em duas ou menos ocasiões no último ano do High School
    - Igual a 1 se o estudante bebeu cinco ou mais bebidas seguidas em três ou mais ocasiões no último ano do High School
    - Igual a NA se a variável indicando quantas vezes o estudante bebeu cinco ou mais bebidas seguidas no último ano do High School for uma NA
- Crie uma variável chamada *free*, que seja:
    - Igual a 0 se o estudante não vive num dormitório onde é proibido consumir bebida alcoólica (*alcohol-free housing*)
    - Igual a 1 se o estudante vive num dormitório onde é proibido consumir bebida alcoólica (*alcohol-free housing*)
    - Igual a NA se a variável indicando se o estudante vive ou não num dormitório onde é proibido consumir bebida alcoólica for NA.
- Crie uma variável chamada *volfree*, que seja:
    - Igual a 1 se *free*=1 (o estudante vive num dormitório onde é proibido ingerir bebida alcoólica) e o aluno requisitou morar num dormitório desse tipo
    - Igual a 0 se *free=1* (o estudante vive num dormitório onde é proibido ingerir bebida alcoólica) e o aluno foi designado a viver num dormitório desse tipo
    - Igual a NA em todos os outros casos
- Crie uma variável chamada *housing*, que seja:
    - Igual a 1 se *free*=0 (o estudante não vive num dormitório onde é proibido ingerir bebida alcoólica)
    - Igual a 2 se *free*=1 (o estudante vive num dormitório onde é proibido ingerir bebida alcoólica) e o aluno foi designado a viver num dormitório desse tipo
    - Igual a 3 se *free* = 1 (o estudante vive num dormitório onde é proibido ingerir bebida alcoólica) e todas as habitações do campus possuem a regra de proibição de consumo de bebida alcoólica.
    - Igual a 4 se *free* = 1 (o estudante vive num dormitório onde é proibido ingerir bebida alcoólica) e o aluno requisitou morar num dormitório desse tipo
    - Igual a NA em todos os outros casos
- Exclua todas as variáveis que não sejam as seis que você acabou de criar.
- Exclua todos os casos onde o valor de *hsdrunk, drunk* ou *housing* seja NA.
- Salve o novo banco de dados na pasta **DadosAnalise/,** nomeando o arquivo de *Analise.RData*.
    - Você salva o banco com a função *save()*.
    - Aqui também você precisará indicar um diretório de trabalho relativo para especificar a localização da pasta

DadosAnalise/.

Lembre-se de salvar tratamento.R na sua pasta ***Scripts/***.

III.B. *ApendiceDados.R*

O *script* ApendiceDados.R irá conter comandos que geram as informações que vão aparecer no Apêndice de Dados (mais informações abaixo).

Criando o *script*

- Abra o RStudio.
- Abra um novo *script* de R.
- Abra o arquivo do banco de análise, analise.RData (que, após ter escrito e rodado tratamento.R, deve estar salvo na pasta **DadosAnalise/**).
- Crie os seguintes resultados para cada uma das cinco variáveis no banco de análise:
    - O número de NAs e o número total de observações (incluindo NAs)
    - A distribuição da frequência e a distribuição da frequência percentual
    - Um gráfico de barra que apresente a distribuição da frequência percentual.

Lembre-se de escrever comentários explicando quais linhas produzem qual informação e gráfico. Cada linha que produzir informação ou gráfico que componha o Apêndice de Dados deve ser precedido por um comentário que descreva o que se espera que esse comando faça. Por exemplo, acima de uma linha que gere uma tabela mostrando a distribuição da frequência da variável *housing*, você escreve um comentário do tipo "A seguinte linha gera uma tabela mostrando a frequência da distribuição de *housing*".

Lembre-se de salvar ApendiceDados.R em sua pasta ***Scripts/***.

III.C. *analise.R*

O analise.R gera os gráficos de barra que constituem a principal análise que você fará para essa atividade, utilizando os dados do seu banco de dados de análise (analise.RData).

Criando o *script*

- Abra o RStudio.
- Abra um novo *script* de R.
- Abra o analise.RData (que, depois de você ter escrito e rodado tratamento.R, deve estar salvo na sua pasta **DadosAnalise/**).
- Crie um gráfico de barras, com duas barras, onde:
    - Uma barra represente estudantes vivendo em dormitórios onde é proibido consumir bebida alcoólica
    - Uma barra represente estudantes que não vivem em dormitórios onde é proibido consumir bebida alcoólica
    - A altura de cada barra é igual à proporção de estudantes (do grupo de estudantes que a barra representa) que beberam o suficiente para ficar bêbado três ou mais vezes nos últimos trinta dias.
    - Salve esse gráfico na pasta **Graficos/** (subpasta de **ExercicioAlcool/**), com o nome Figura1.jpg. (Use um diretório de trabalho relativo para especificar a localização da subpasta **Graficos/**)
- Crie um gráfico de barra, com duas barras, onde:
    - Uma barra represente estudantes que

vivem em dormitórios onde é proibido consumir bebida alcoólica porque eles requisitaram isso
    - Uma barra represente estudantes que vivem em dormitórios onde é proibido consumir bebida alcoólica porque eles foram designados
    - A altura de cada barra deve ser igual à proporção de estudantes (no grupo representado pelas barras) que beberam o suficiente para ficar bêbados três ou mais vezes nos últimos trinta dias
    - Salve esse gráfico na pasta **Graficos/** (subpasta da pasta **ExercicioAlcool/**), com o nome Figura2.jpg.
- Crie um gráfico de barrar, com quatro barras, onde:
    - Cada barra represente estudantes em uma das quatro categorias definidas na variável *housing*
    - A altura de cada barra é igual à proporção de estudantes (no grupo representado pelas barras) que beberam o suficiente para ficar bêbados três ou mais vezes nos últimos trinta dias
    - Salve esse gráfico na pasta **Graficos/** (subpasta da pasta **ExercicioAlcool/**), com o nome Figura3.jpg.
- Crie dois gráficos de barra (lado-a-lado), cada um possuindo duas barras, onde:
    - Um dos gráficos represente apenas estudantes que tiveram cinco ou mais *drinks* em sequência em três ou mais ocasiões no último ano do *High School*
    - Um dos gráficos represente apenas estudantes que tiveram cinco ou mais *drinks* em sequência em menos de três ocasiões em seu último ano de *High School*
    - E em cada um desses gráficos,
        - Uma barra representa estudantes que vivem em dormitórios onde o consumo de álcool é proibido
        - Uma barra representa estudantes que não vivem em dormitórios onde há essa proibição
        - A altura de cada barra é igual à proporção de estudantes (no grupo representado pelas barras) que beberam o suficiente para ficar bêbados três ou mais vezes nos últimos trinta dias
        - Salve esse gráfico na pasta **Graficos/** (subpasta da pasta **ExercicioAlcool/**), com o nome Figura4.jpg.
- Crie dois gráficos de barra (lado-a-lado), cada um possuindo duas barras, onde:
    - Um dos gráficos represente apenas estudantes que tiveram cinco ou mais *drinks* em sequência em três ou mais ocasiões no último ano do *High School*
    - Um dos gráficos represente apenas estudantes que tiveram cinco ou mais *drinks* em sequência em menos de três ocasiões em seu último ano de *High School*
    - E em cada um desses gráficos,
        - Uma barra represente estudantes que vivem em dormitórios onde é proibido o consumo de bebidas alcoólicas porque requisitaram
        - Uma barra represente estudantes que vivem em dormitórios onde é proibido o consumo de bebida alcoólica por que eles foram designa-

dos
    - A altura de cada barra é igual à proporção de estudantes (no grupo representado pelas barras) que beberam o suficiente para ficar bêbados três ou mais vezes nos últimos trinta dias
    - Salve esse gráfico na pasta **Graficos/** (subpasta da pasta **ExercicioAlcool/**), com o nome Figura5.jpg.
- Crie dois gráficos de barra (lado-a-lado), cada um possuindo quatro barras, onde:
    - Um dos gráficos represente apenas estudantes que tiveram cinco ou mais *drinks* em sequência em três ou mais ocasiões no último ano do *High School*
    - Um dos gráficos represente apenas estudantes que tiveram cinco ou mais *drinks* em sequência em menos de três ocasiões em seu último ano do *High School*
    - E em cada um desses gráficos,
        - Cada barra represente estudantes em uma das quatro categorias definidas na variável *housing*
        - A altura de cada barra é igual à proporção de estudantes (no grupo representado pelas barras) que beberam o suficiente para ficar bêbados três ou mais vezes nos últimos trinta dias
        - Salve esse gráfico na pasta **Graficos/** (subpasta da pasta **ExercicioAlcool/**), com o nome Figura6.jpg.

Escreva comentários indicando quais linhas produziram que gráficos. Por exemplo, antes da linha que gera a Figura 4, escreva um comentário como "A seguinte linha de comando gera a Figura 4".

Lembre-se de salvar analise.R em sua pasta de ***Scripts&Relatorio/***.

IV. O RELATÓRIO

O relatório que você deve entregar para essa atividade consistirá na apresentação das Figuras que foram geradas e de respostas a uma série de perguntas que pedem que você interprete os gráficos de barra que você criou. Salve em *.pdf.*

Apresente as Figuras e escreva uma legenda abaixo de cada gráfico que indique a numeração e um título informativo. Por exemplo, uma legenda apropriada para a Figura 1 seria: "Figura 1: Taxa de Consumo Excessivo de Álcool em Dormitórios onde é proibido consumir álcool vs. Dormitórios onde não há essa proibição".

Questões

1) Descreva a amostra no arquivo BancoAnalise.dta que você criou. Qual a unidade de análise? Quantas observações existem? Descreva o método utilizado para criar a amostra e quais critérios foram considerados para incluir ou excluir grupos de indivíduos.

Note que essa pergunta é sobre seu banco limpo e processado, ao invés do banco original que você baixou do ICPSR. De qualquer forma, para responder, você terá que recorrer ao *codebook* do banco original.

2) Explique em suas palavras o que a Figura 1 apresenta. Era o que você esperava? Se não, como a Figura 1 é diferente do que você esperava?

3) Explique em suas palavras o que a Figura 2 apresenta. O que você vê na Figura 2 repercute de alguma forma o

que você viu na Figura 1? Explique.

4) Quantos indivíduos são representados na Figura 1? (Ou seja, quantas observações foram utilizadas para gerar a figura?). E quantos indivíduos estão representados na Figura 2? Se o número de indivíduos representados na Figura 1 e Figura 2 não é o mesmo, explique: alguns indivíduos apresentados na Figura 1 não estão representados na Figura 2? Se sim, que indivíduos foram esse? E alguns indivíduos apresentados na Figura 2 não foram representados na Figura 1? Se sim, quais?

5) Considerando a Figura 3, quais partes contêm informação que já foram apresentadas nas Figuras 1 e 2? Que partes apresentam nova informação? Resuma em suas palavras o que a Figura 3 apresenta.

6) Compare a Figura 4 com a Figura 1. Explique que novas informações são fornecidas pela Figura 4 e apresente uma interpretação dessas informações.

7) Compare a Figura 5 com a Figura 2. Explique que nova informação é fornecida pela Figura 4 e apresente uma interpretação.

8) Compare a Figura 6 com a Figura 3. Explique que nova informação é fornecida pela Figura 4 e apresente uma interpretação.

9) Das figuras que você criou, quais conclusões gerais podem ser retiradas sobre os fatores associados ao consumo de álcool entre estudantes universitários?

V. O APÊNDICE DE DADOS

O Apêndice de Dados deve consistir em múltiplas seções, uma para cada variável do banco de dados de análise. Salve também em *.pdf*.

Para cada variável, a seção deve fornecer:

- O nome da variável.
- O número de NAs no formato NA/Total (quantidade de NAs/número total de observações).
- Uma definição da variável.
- Possíveis valores da variável.
- Uma explicação do que cada valor da variável representa (codificação dos valores).
- Uma tabela apresentando a distribuição da frequência e a distribuição da frequência percentual. Se a variável categórica é ordenada, essa tabela deve também apresentar distribuição da frequência percentual acumulativa.
- Um gráfico de barra que ilustre a distribuição da frequência percentual.

VI. O ARQUIVO LEIA-ME

O arquivo Leia-Me é um documento que apresenta informações sobre os arquivos eletrônicos que você organizou para documentar o trabalho desta análise. Deve consistir em duas seções principais:

> A primeira seção deve apresentar um mapa de todos os arquivos incluídos na documentação da replicação, assim como as pastas e subpastas onde estão localizados.
>
> A segunda seção deve fornecer instruções passo-a-passo explicando como utilizar a documentação para (i) replicar todo o processamento de dados necessário para transformar o banco original no banco de análise, (ii) gerar todos os resultados que você apresenta no Apêndice de Dados, e (iii) reproduzir os seis gráficos de barra que você criou.

Essas instruções devem ser escritas em português claro e devem ser detalhadas o suficiente para que alguém não familiarizado com esse exercício consiga seguir e reproduzir tudo que você fez. Salve em *.pdf*.

VII. DOCUMENTAÇÃO ELETRÔNICA

Uma vez finalizado o exercício, os arquivos utilizados e criados devem estar alocados na hierarquia de pastas da seguinte maneira:

ExercicioAlcool/
Leia-Me.pdf
Relatorio.pdf
Dados/ (subpasta de **ExercicioAlcool/**)
DadosOriginais (subpasta de **Dados/**)
04291-0001-Data.dta
04291-0001-Codebook.pdf
DadosAnalise (subpasta e **Dados/**)
BancoAnalise.dta
Scripts/ (subpasta de **ExercicioAlcool/**)
tratamento.R
ApendiceDados.R
analise.R
Graficos/ (subpasta de **ExercicioAlcool/**)
Figura1.jpg
Figura2. jpg
Figura3. jpg
Figura4. jpg
Figura5. jpg
Figura6. jpg

Não deixe nada além disso nessas pastas

- Sem nenhuma pasta além dessas.
- Sem nenhum arquivo além desses

Teste seus arquivos para saber se rodam

- Encontre um computador que não seja o que você realizou a atividade e copie para ele a pasta inteira, com todo seu conteúdo.
- Siga os seguintes passos, sem modificar a ordem das pastas ou os arquivos:
  - Inicie o R;
  - Defina ***Scripts/*** como seu working directory
  - Rode o tratamento.R e após isso, confira se o arquivo BancoAnalise.dta que você tem na pasta **DadosAnalise/** foi alterado por uma nova cópia criada recentemente (confira a data e hora de modificação do arquivo na pasta)
  - Rode o ApendiceDados.R e verifique se gera os resultados esperados. Confira também se os gráficos na pasta **Graficos/** foram atualizados por novas versões recentes.

Rode o analise.R e confira se os seis arquivos de gráficos localizados na pasta **Graficos/** foram atualizados.

6. Outras atividades do TIER

Com o intuito de promover a importância da transparência e de difundir os valores da ética na ciência, o Projeto TIER desenvolve algumas atividades e eventos de promoção de seu protocolo e de boas práticas científicas. Todas podem ser encontradas em maiores detalhes em seu endereço da web[21].O Projeto TIER promove workshops voltados a professores e pesquisadores universitários, tanto para ouvi-los sobre suas práticas em prol da transparência e da replicabilidade na ciência, quanto para apresentá-los às ideias mais recentes sobre o protocolo (e colher comentários e impressões)[22]. Esses workshops são ótimas oportunidades de cooperação interuniversidades e de cooperação internacional, em termos de boas práticas científicas, reunindo professores de distintos campos das ciências sociais e diferentes países.

Há ainda o programa de *Fellowship,* voltado a jovens professores pesquisadores engajados com ensino e treinamento de métodos quantitativos sob uma lente

21 Ver: https://www.projecttier.org/ . Acesso em 20 de março de 2020.

22 Ver: https://www.projecttier.org/fellowships-and-workshops/fall-2019-faculty-development-workshop/ . Acesso em 20 de março de 2020.

transparente e replicável. As bolsas valem por um ano acadêmico e buscam incentivar que o recipiente desenvolva e difunda métodos de ensino de pesquisa transparente – como exercícios *soup-to-nuts* e *workshops*. Além disso, o TIER promoveu também nos meses de fevereiro e março de 2020 uma série de conferências no formato de webcast. A cada semana, um convidado fazia uma fala sobre os seus mais recentes desenvolvimentos e ideias sobre transparência e replicabilidade. Dentre esses "líderes da transparência", estavam Gary King, Dorothy Bishop e Scott Long. As apresentações estão disponíveis no site do TIER[23].

7. Considerações Finais

Os princípios e as práticas de transparência têm se tornado o padrão-ouro da ciência empírica. Iniciativas para difundir essas ideias e apresentar ferramentas para aumentar a transparência nas pesquisas e na produção científica sobre o assunto tem se tornado cada vez mais relevantes. Materiais de replicação ganham força enquanto elementos que compõe uma pesquisa concluída, e suas divulgações e compartilhamentos têm sido um requisito para publicação em diversos periódicos especializados, seja em território americano e, embora ainda de forma incipiente, aqui no Brasil.

Nesse trabalho, tivemos o objetivo de discutir formas de aumentar a transparência de pesquisas empíricas nas Ciências Sociais, principalmente através da documentação dos materiais de replicação. Apresentamos, ao longo do artigo, o surgimento do Protocolo TIER e seu enquadramento nas três dimensões do padrão de replicação (substantiva, pedagógica e transparência), uma vez que o protocolo permite a continuidade da produção acadêmica e a movimentação da roda da ciência, possibilita que jovens pesquisadoras sejam iniciadas no mundo na pesquisa empírica através de dados estatísticos e análise de dados, além de aumentar o nível de transparência dos trabalhos que o adotam. Além disso, apresentamos a estrutura organizacional do protocolo, os arquivos comportados e a hierarquia de pastas que pressupõem a documentação sugerida pelos criadores. Apresentamos, ainda, uma versão encurtada de um exercício *soup-to-nuts* para implementação do protocolo, ideal para aplicação em turmas de introdução a estatística e análise de dados. Esperamos que os alunos que sejam introduzidos a essas análises já dentro de um paradigma transparente e replicável tenham não só mais facilidade em aprender e em desenvolvê-las, mas que também carreguem consigo esses valores de boa ciência.

Com isso, visamos indicar um caminho menos tortuoso para que a pesquisadora alcance níveis mais altos de transparência em sua pesquisa. Além disso, esperamos ter difundido e apresentado de forma simples e didática o protocolo. Dessa maneira, esperamos contribuir com a agenda de pesquisa sobre transparência e replicabilidade, que vem crescendo na Ciência Social como um todo, e na Ciência Política brasileira em especial.

23 Ver: https://www.projecttier.org/fellowships-and-workshops/weekly-webcast-leaders-research-transparency/ . Acesso em 20 de março de 2020.

REFERÊNCIAS BIBLIOGRÁFICAS

ALVAREZ, R. M; KEY, E. M; NÚÑEZ, L. (2018). "Research replication: Practical considerations". *PS: Political Science & Politics,* 51(2), 422-426.

BALL, R.; MEDEIROS, N. (2012). "Teaching Integrity in Empirical Research: A Protocol for Documenting Data Management and Analysis". *The Journal of Economic Education*, 43(2), 182-189.

BALL, R. (2018). Animal House in Alcohol-Free Dorms? TIER Project. Disponível em: https://www.projecttier.org/tier-classroom/soup-nuts-exercises/#shorter-exercises-for-teaching-transparency-and-reproducibility. Acesso em 06 de outubro de 2020.

CASEY, K; GLENNERSTER, R; MIGUEL, E. (2012). "Reshaping institutions: Evidence on aid impacts using a preanalysis plan". *The Quarterly Journal of Economics*, 127(4), 1755-1812.

CHRISTENSEN, G.; SODERBERG, C. (2016). "Manual of best practices in transparent social science research". *Retrieved April*, 2, 2017.

CHRISTENSEN, G.; MIGUEL, E. (2018). "Transparency, reproducibility, and the credibility of economics research". *Journal of Economic Literature*, 56(3), 920-80.

DE LA GUARDIA, F. H.; STURDY, J. (2019). "Best Practices for Transparent, Reproducible, and Ethical Research". *IDB Technical Note* ; 1635.

FIGUEIREDO FILHO, D.; LINS, R.; DOMINGOS, A.; JANZ, N.; SILVA, L. (2019). "Seven Reasons Why: A User's Guide to Transparency and Reproducibility". *Brazilian Political Science Review*, 13(2).

JANZ, N. (2016). "Bringing the gold standard into the classroom: replication in university teaching". *International Studies Perspectives*, 17(4), 392-407.

KEY, Ellen M. (2016). "How are we doing? Data access and replication in political science. PS: Political Science & Politics, 2016, 49.2: 268-272

KING, G. (1995). "Replication, replication". *PS: Political Science & Politics*, 28(3), 444-452.

MIGUEL, E., CAMERER, C., CASEY, K., COHEN, J., ESTERLING, K. M., GERBER, A., LAITIN, D. (2014). "Promoting transparency in social science research". *Science*, 343(6166), 30-31.

MUNAFÒ, M. R., NOSEK, B. A., BISHOP, D. V., BUTTON, K. S., CHAMBERS, C. D., DU SERT, N. P., IOANNIDIS, J. P. (2017). "A manifesto for reproducible science". *Nature human behaviour*, 1(1), 1-9.

OLKEN, B. A. (2015). "Promises and perils of pre-analysis plans". *Journal of Economic Perspectives,* 29(3), 61-80.

OSINSKI, L. (2019). "PROOF course Open Science: The new default in science, 01-10-2019". *Presentation at Eindhoven University of Technology*. Disponível em: <https://www.projecttier.org/tier-classroom/course-materials/#modal230>. Acesso em 23 mar. 2020.

PARANHOS, R., FIGUEIREDO FILHO, D. B., da Rocha, E. C., da Silva Jr, J. A., & SANTOS, M. L. W. D. (2012). "Levando Gary King a sério: desenhos de pesquisa em Ciência Política". *Revista Eletrônica de Ciência Política*, 3(1-2).

PARANHOS, R., FIGUEIREDO FILHO, D. B., da ROCHA, E. C., CARMO, E. F. (2013). "A importância da replicabilidade na ciência política: o caso do SIGOBR". *Revista Política Hoje*, 22(2), 213-229.

PRZEWORSKI, A., SALOMON, F. (1988). "The Art of Writing Proposals: Some candid suggestions for Applicants to Social Science Research Council". *SSRC: New York*.

A map for transparency and replicability in empirical social science: the TIER Protocol

Amanda Domingos - Federal University of Pernambuco
Ian Rebouças Batista- Federal University of Pernambuco

Resumo

Transparência e replicabilidade são o padrão ouro da pesquisa científica. Mas, alcançar esse padrão pode ser um caminho longo e trabalhoso. Nesse artigo, apresentamos uma maneira de aumentar a transparência de sua pesquisa através da documentação. Argumentamos que o Protocolo TIER é uma importante ferramenta de transparência e de auxílio na organização de materiais para replicação. Trata-se de uma metodologia de documentação que compila os principais materiais produzidos numa pesquisa empírica social. Além disso, destacamos a adequação do protocolo nas três dimensões do padrão de replicação (substantiva, pedagógica e transparente), uma vez que contribui para o avanço da ciência, introduz jovens graduandas à análise de dados de uma forma prática e permite a transparência quanto aos procedimentos empregados na pesquisa. Por fim, discutimos atividades relacionadas ao Projeto TIER que contribuem com os objetivos do protocolo. Especificamente, apresentamos um exercício de introdução à análise de dados a ser realizado no R, onde a documentação do protocolo é requerida e sua aplicação, portanto, ilustrada.

Palavras-chave: Transparência; Replicabilidade; Documentação; Protocolo TIER

Abstract

Transparency and replicability are the gold standard in empirical research. But reaching that standard can be a long and tortuous journey. In this article, we present a way to increase the transparency of research through documentation. We argue that the TIER Protocol is an essential tool for transparency and assistance in organizing replication materials. It is a documentation methodology that compiles the main materials produced in empirical social research. Also, we highlight the adequacy of the protocol in the three dimensions of the replication standard (substantive, pedagogical, and transparent) – since it contributes to the advancement of science, introduces young undergraduate students to data analysis practically, and allows transparency of research procedures. Finally, we discuss activities related to the TIER Project that contribute to the objectives of the protocol. Specifically, we present a data analysis introduction exercise for R, where the protocol documentation is required and its application, therefore, exemplified.

Keywords: Replicability; Documentation; TIER Protocol

1. Introduction

Download the dataset, clean it, write scripts, program functions, execute commands and analyze results. If these are routine steps in your research, a protocol will help you organize and publish your results in a transparent and replicable way: the TIER Protocol.

Principles and practices of transparency and replicability have drawn attention in the scientific community, both in national (PARANHOS, 2012; PARANHOS et al., 2014; FIGUEIREDO et al., 2019; AVELINO and DESPOSATO, 2018) and international publications (CHRISTENSEN, 2016; MUNAFÓ et al., 2017; CHRISTENSEN and MIGUEL, 2018; STOCKEMER, KOEHLER E LENZ, 2018; DE LA GUARDIA and STUDY, 2019). Also, such practices have been encouraged through inter-university initiatives such as the *Berkeley Initiative for Transparency in the Social Sciences* (BITSS) and the *Teaching Integrity in Empirical Research* (TIER), which promote courses, workshops, and events focused on the exchange of experiences and ideas in favor of transparency.

Here we understand transparency regarding the procedures adopted in empirical research, such as the availability of the materials used and the indication of the steps taken until reaching the results. In addition to the benefits – such as allowing public access to methodological choices and enabling that the work is replicated and more rigorously evaluated –, transparency also allows: (i) more accurate hypothesis testing, (ii) collaboration with other researchers with similar interests, and (iii) increase in the author's credibility. We argue that a transparent researcher, who documents their work as it is done, has an easier process in elaborating her research. Rigorous documentation brings benefits in terms of scientific reputation, given that it contributes to good practices in science and generates practical benefits, such as increased citations (FIGUEIREDO et al., 2019).

Given this, some instruments have been used to disseminate and implement transparency and replicability in the Social Sciences, with three leading practices: disclosure plan, preregistration, and open data and materials (MIGUEL et al., 2014). This study aims to focus on the third kind. In this paper, we present and suggest adopting the TIER Protocol, a methodology for the documentation of the files used throughout the development of the research (dataset, scripts, outputs), and later accessibility of these materials for replication. We believe that adopting a specific rule regarding the organization of research files increases transparency and favors a study's replicability.

The paper is divided as follows. In the next section, we present the gold-standard of scientific research regarding the transparency of procedures and replicability of results and point out why the reader should aim towards these values and practices. Next, we present the TIER Protocol, highlighting its essential aspects. Then, we emphasize the benefits and incentives of adopting a documentation and disclosure strategy to make research transparent. Lastly, we list some activities developed by the TIER project that corroborate and seek to disseminate good practices in transparency and replicability in the Social Sciences and present an introductory exercise to the protocol. The conclusion wraps up the paper.

2. Why a transparency protocol?

Among the characteristics of good scientific research are "conceptual innovation, methodological rigor, and rich, substantive content" (PRZEWORSKI and SALOMON, 1998, p. 1). In the Social Sciences, however,

two other elements have been added to these over the past few decades: transparency and replicability, held as the gold-standard of empirical research (KING, 1995; PARANHOS et al., 2014; JANZ, 2016). There are many benefits of transparency both for the discipline as a whole and the researcher in particular. Adoption of transparent practices allows the wheel of science to spin faster by allowing more precise hypothesis testing, facilitating collaboration with other researchers with similar interests, and increasing the credibility of the research, as well as removing the usual attention given to the statistical significance and increasing the relevance of research *per se* (DE LA GUARDIA and STUDY, 2019).

Regarding benefits to authors and journals, Figueiredo et al. (2019) present seven reasons to adopt transparent procedures in empirical research in the Social Sciences: (I) to prevent honest mistakes or even frauds in empirical analyses; (II) to making writing scientific papers easier, since it allows testing other hypotheses with the same dataset; (III) transparent procedures allow reviewers to have access to data and scripts and therefore enables them to give more precise recommendations on how to improve the paper; (IV) by ensuring systematization of raw and analyzed data, it enables the author's scientific work to continue; (V) it contributes to data analysis learning, by allowing students to have contact with data and statistical procedures used in real life; and (VII) it increases the impact of academic work since empirical evidence shows that papers that adopts transparent procedures and practices are cited twice as often as those who are not transparent.

On the other hand, there are many costs in not adopting transparency and replicability procedures. Studies that do not maintain public procedures may contribute to lowering confidence in their results, and it may be an obstacle to publishing in high-impact factor journals that request replication materials – diminishing the study's range (MUNAFO et al., 2017; DE LA GUARDIA and STUDY, 2019). Thus, someone could argue that adopting transparency principles and practices are too much work. But it is worth remembering that even the gold-standard of replication does not require that an individual reproduce the results presented in books or scientific papers, only recommends that is the author makes available as much information as needed so that a researcher who desires to can reproduce them (KING, 1995).

Despite that, the practice of transparency is not the rule. Roughly 67% of papers published in the *American Political Science Review* between 2013 and 2014 are not replicable (KEY, 2016). In Brazil, only 5% of papers published between 2012 and 2016 in Social Sciences journals were apt for complete replication (AVELINO e DESPOSATO, 2018)[12]. The standard is still obscurity, not transparency. Consequently, it is not unusual to find researchers who have not made public the procedures and practices that led them to their scientific products' conclusions. We ask the dear reader: how many times, in the production of your empirical research, did you make an effort to remember the paths that took you to the result you found? More than that, what have you done to be transparent? If you thought a bit about that, this article could help you. Over the next sections, we will suggest an organization protocol to facilitate transparent practices and discuss the documentation process of empirical research.

3. The TIER Protocol

The TIER Protocol is a documentation methodology for replication materials in empirical research. Simply put, the protocol gives a general indication for the

1 The report was a result of the project "*Research Transparency in Brazilian Political and Social Science: A First Look*", by the authors along with the *Berkeley Initiative for Transparency in the Social Sciences* (BITSS). For more information, visit: < https://www.bitss.org/projects/research-transparency-in-brazilian-political-and-social-science-a-first-look/>. Last acessed in: 20 March 2010.

2 The final report is publicly avaliable at: <https://osf.io/f279z/>. Last acessed in 20 March 2020.

organization of files used in a study, the construction of folders, and what information must be made available in each of them. Initially, the protocol was developed to deal with a problem in the classroom. The papers handed in by the students of Introduction to Statistics, discipline taught by Richard Ball for economics undergraduate students at Haverford College (USA), had a similar issue to those in scientific papers published in great academic journals: they were not transparent. When handing in the discipline's final papers with wrong results, it was practically impossible for the professor to understand at what point in the data analysis the procedures were incorrect conducted, which hindered adequate grading by the professor and, consequently, the student's learning process.

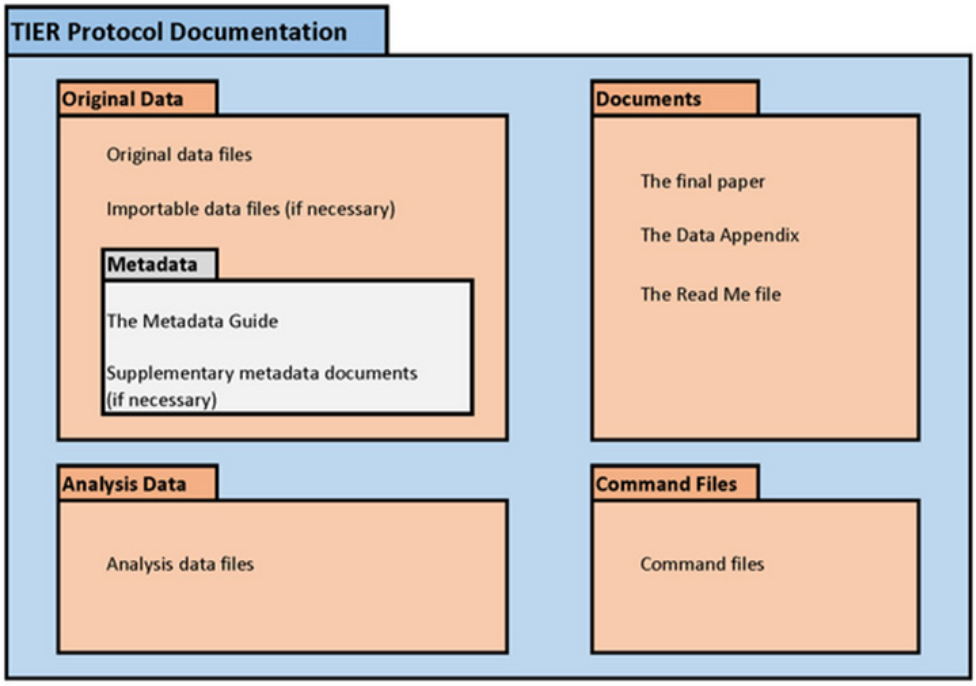
To guarantee a standard in the final class project delivery, enabling the professor to grade efficiently (having access to scripts and data in an intelligible and organized way), Ball elaborated, with the College's librarian, Norm Medeiros, the TIER Protocol (BALL and MEDEIROS, 2012). The protocol consists of a specific organization and hierarchy of folders that separate information on a given research (raw data, analyzed data, scripts, etc.) and facilitates the replication of the findings. Despite its pedagogical origin, the TIER Protocol goes beyond the classroom. Several researchers in the field of Social Sciences can use the protocol to increase the level of transparency in their research. Thus, we argue that adopting the TIER Protocol results in the collaboration of the three dimensions in the replication standard: substantive, pedagogical, and transparency (PARANHOS et al., 2013).

The TIER Protocol has a general guideline, as illustrated in Figure 1, and has versions that improve the organization and operationalization of the work's replication, as with version 4.0 in Figure 2. We will discuss the two arrangements next, but some procedural clarifications of TIER's organizational methodology purposes help in understanding.

The folders that will comprise the work's replication material must be created at the beginning of the research. They will stay with the researcher throughout the whole process: treatment of the original dataset, running analyses, writing the paper and appendices. The researcher must save and work with the files in the proper folders and, in their scripts or command file, referencing the working directories according to the folder hierarchy, making it essential that the folder hierarchy is maintained throughout the project. For instance, in the script where you treat the original dataset and create the set for analysis, you must load it from the folder *OriginalData* and save the treated dataset in the folder *AnalysisData* (Figure 1). The idea is that when using software such as R, the script has the lines needed both to open the original data and also save the analysis data in the correct folders[3].

3 As a replicability measure, ideally one would adopt a relative working directory. For more information, see: https://ssc.wisc.edu/sscc/pubs/R_intro/book/1-10-paths-and-working-directories.html. Last accessed on: 07 October 2020.

Figure 1: Overall view of the documentation suggested by the TIER Protocol



Source: Project TIER.4

But what folder organization should you adopt since the structures proposed in Figures 1 and 2 are different? The protocol's main goal is to provide a guideline for organizing and making accessible all digital material used during empirical research. It is important that this guideline is coherent and , if not intuitive, justified and presented. Therefore, the folders' exact configuration and files are up to you. Figure 1 is the most generic example, from which four other versions were elaborated, and Figure 2 is the most recent formatting suggested by TIER Project (4.0). Therefore, once the protocol's argument on research documentation and transparency convinces the researcher/reader, she is encouraged to organize her material based on a protocol that works for her – that is, in the hierarchy of folders that she finds convenient.

For individual protocols to work for any future replicator, a few basic guidelines that compound TIER's message are required. The first one is creating ReadMe files (usually in .txt, to facilitate reading regardless of the operational system), which will guide the replication activity from the beginning. One of the premises of a transparent replication is that this kind of files are the first to be read because, in addition to presenting and explaining the adopted folder configuration that the replicator will find when reading, these are the text documents that say which files to read first and what will be the sequence of the replication work. Therefore, it is useful that it appears in the first folder, as shown in Figure 2. The elaboration of a clear and direct ReadMe file shows the researcher's concern with the future replication of their findings. Thus, it must offer a complete overview of how the folders are organized and replication steps.

For instance, for a paper using the protocol inspired by what is shown in Figure 2, the ReadMe file might contain a text similar to this: "Initially, go to the *Treatment* subfolder, located in the *Scripts* folder, and open the

4 Original template used by the TIER Project. Available at: https://www.projecttier.org/tier-protocol/specifications-3-0/#overview-of-the-documentation. Last accessed on 22 January 2021.

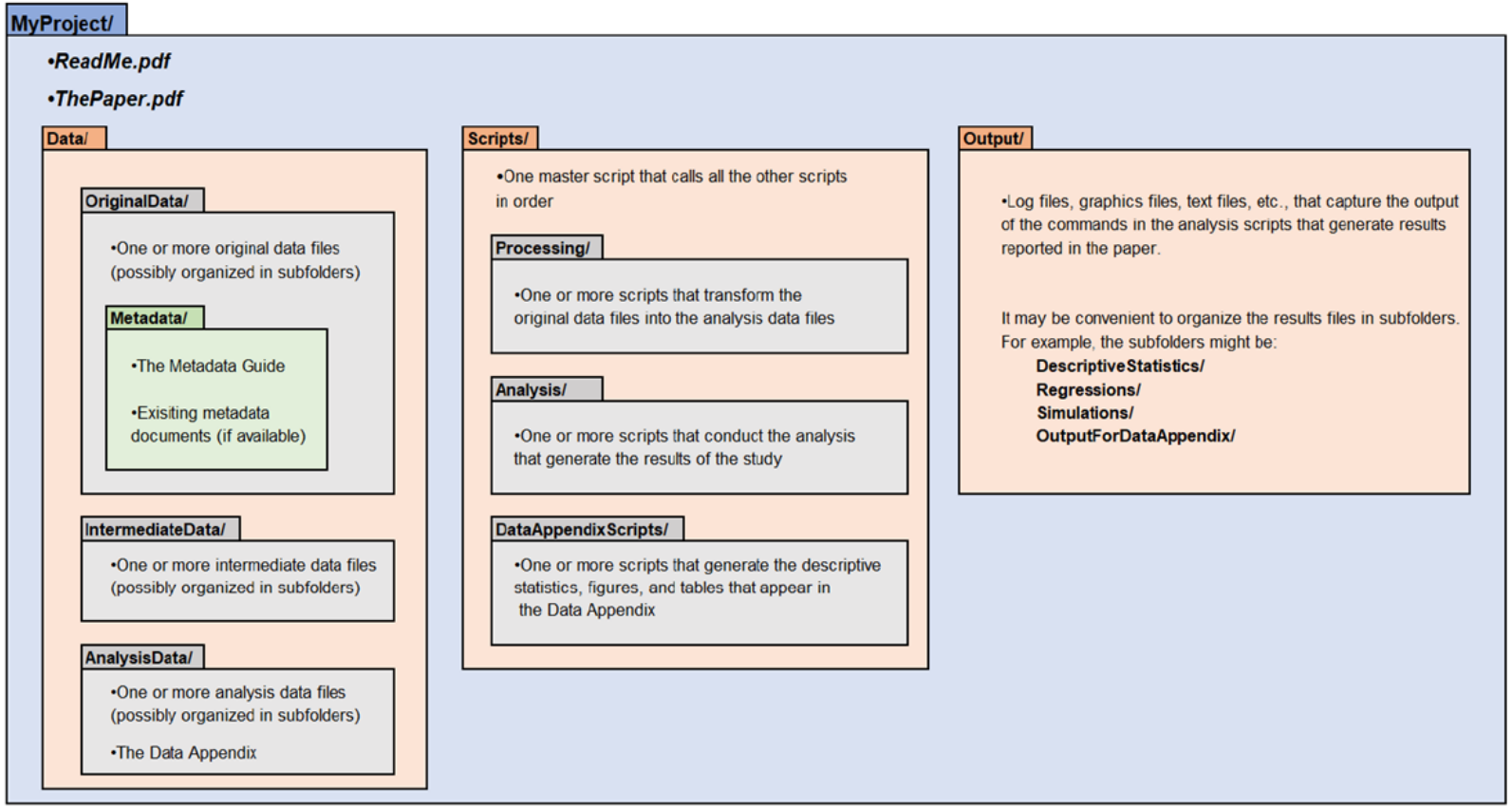
Treatment.R script. With it, you will treat Database. dta, which is the *OriginalData* subfolder, within the *Data* folder, and will generate NewDatabase.dta, which will be saved in the subfolder *AnalysisData*. Next, also in the *Scripts* folder, go to the *Analysis* subfolder and open the Analysis.R script. With it, you will generate Graphs 1, 2, and 3 and you will save it in the *Results* folder". ReadMe files can also contain information on the software used for data analysis, web addresses where the original data were downloaded, or even the researcher's contact information.

Like ReadMe, but with a different function, the TIER Protocol considers it fundamental that the researcher creates a Data Appendix (or codebook). In this document, the variables present in the analysis dataset are described in detail, including their type, how they were measured, and how they were built (if it was already in the original dataset or if it was manipulated by the researcher while treating the data). Descriptive statistics on each variable can also be included. Also, in this document, the research gives more in-depth information on the software's version and which statistical packages were used.

Another TIER Protocol's basic guideline on any of its versions, and that must also be present in a possible personal protocol inspired by it, is the separation of files by types. In both Figures 1 and 2, there are folders for datasets, scripts, and documents. We will analyze these figures in more detail now.

Figure 1, the protocol's general guideline, has four folders in total: *OriginalData; AnalysisData; CommandFiles; Documents*. Within the *OriginalData* folder, we find the original dataset, directly downloaded from whatever source, the converted dataset, should the researcher have had to convert the original downloaded dataset (note that, even then, it is crucial to keep the dataset in the original format), and a subfolder called *Metadata*, where a Metadata Guide (a Data Appendix for the original dataset, including the bibliographical reference, the DOI identifier, the date of its download, and the web address), and supplementary documents from this original dataset (if there should be any). In the *AnalysisData* folder, there is the treated dataset(s), the one with the researcher will conduct the analysis. We found the final paper, the Data Appendix, and the ReadMe file in the Documents folder. Lastly, in the *CommandFiles* folder, there will be all the scripts used in the research, from the treatment of the original dataset up to the analyses carried out.

Figure 2: TIER Protocol version 4.0 (2020)



Source: Project TIER. Tradução dos autores

Figure 2 is based on a preliminary elaboration of the protocol's version 4.0, presented in November 2019. In the main folder, there are immediately two files, the final paper, the ReadMe file, and three folders: *Data*, *Scripts*, and *Results*. There are three subfolders in the Data folder: *OriginalData* (with the *Metadata* folder contained within); *IntermediateData*, where there will be datasets that might have been converted to a different format; and *AnalysisData*, where the *DataAppendix* will also be present. In the *Scripts* folder, there is a "master" script, which combines all the work's scripts and three subfolders: *Treatment*, *Analysis*, and *DataAppendixScript*, each folder containing the script for the corresponding step. Lastly, in the *Results* folder, all the files that were created throughout the analysis, such as tables, graphs, and figures, which can also be subdivided into folders.

Both versions of the protocol share the same ideal of research transparency and documentation, although adopting different folder organizations and hierarchies. Understanding the protocol's goals and the logic behind documentation, the researcher interested in transparency and replicability can, if she does not adopt one of the main versions, create her own TIER Protocol version. A version that makes sense to her and benefits her workflow while also being clear and accessible to anyone interested in replicating their research results.

After pointing out all the material following the TIER Protocol, availability must occur in the most accessible way possible. Since data transparency is a growing issue, we have at our disposal a range of digital resources that offer an adequate structure. The *Open Science Framework*[5] (OSF), *Dataverse*[6], and *Github*[7] are three examples of suitable repositories for sharing all work material, including the possibility of subdividing the folders of the protocol. In this paper, we consider a quantitative approach, given that the nature of the protocol benefits this type of work. But there are also digital repositories for papers with qualitative methods[8], such as the *Qualitative Data Repository*[9], the UK data archive[10], and the National Anthropological Archives[11].

4. Documentation, TIER Protocol, and the dimensions of the replication standard

In 1995, the year the "*verification/replication*"[12] issue was released, the reality of scientific production in the Social Sciences was different. The dossier proposed to researchers to make available more details on the empirical choices conducted in their research, in footnotes and methodological appendices. But these efforts were met with the limits imposed by journals and editors in publishing these details. On the other hand, availability and storage of information were a problem: the facility of hard drives and clouds did not exist; the digital archive related to publications from the *Inter-University Consortium for Political and Social Research* (ICPSR)[13], for example, would only be released sometime after the issue (ALVAREZ et al., 2018).

With the development of suitable online platforms and repositories, information disclosure has become eas-

5 See: https://osf.io . Last accessed on 20 March 2020.

6 See: https://dataverse.org/ . Last accessed on 20 March 2020.

7 See: https://github.com/ . Last accessed on 20 March 2020.

8 For more information on repositories that accept qualitative data, see: https://www.sesync.org/list-of-qualitative-data-repositories. Last accessed on 24 March 2020.

9 See: https://qdr.syr.edu/. Last accessed on 24 march 2020.

10 See: https://www.ukdataservice.ac.uk/manage-data. Last accessed on 24 March 2020.

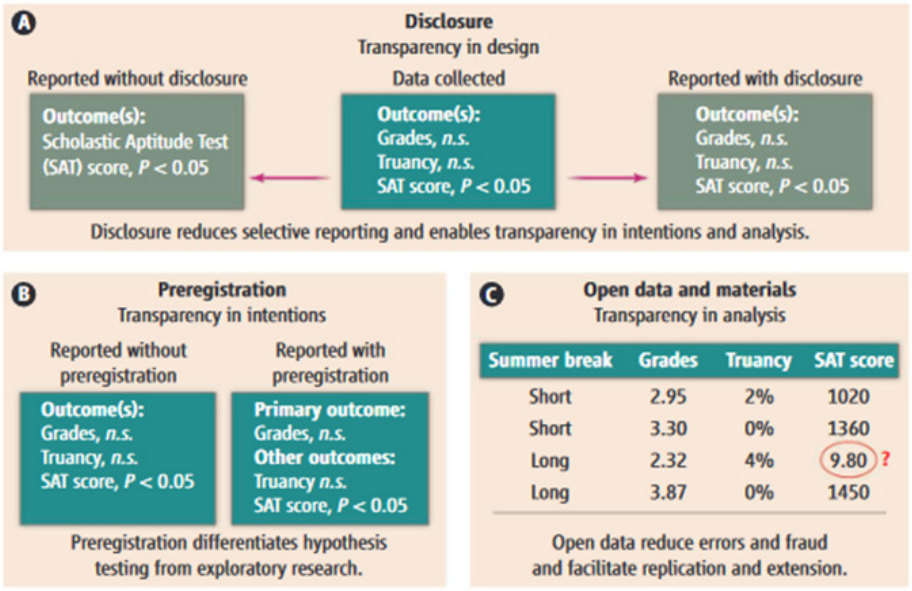
11 See: http://anthropology.si.edu/naa/search.html. Last accessed on 24 March 2020.

12 The dossier is available at: < https://www.cambridge.org/core/journals/ps-political-science-and-politics/issue/DAEAAA8412CD56791EAFCC75082A4C15>. Last accessed on 24 March 2020.

13 Available at: <www.icpsr.umich.edu/icpsrweb/deposit/pra/index.jsp>. Last accessed on 24 March 2020.

ier every day. Consequently, several efforts have been undertaken to disseminate and implement the principles and practices of transparency and replicability in the Social Sciences. These efforts have been concentrated on three leading practices: disclosure plans, preregistration, and open data and materials (MIGUEL et al., 2014), as illustrated in Figure 3.

Figure 3: Three mechanisms to increase transparency in scientific reports (Demonstration with the research question "do shorter summer breaks improve educational outcomes?", where "n.s" means P > 0,05.)



Source: Miguel et al. (2014), p.30.

Disclosure plans arise to argue for the production of systematic reports that seek to ensure that researchers document and disclose the way they proceeded regarding data collection and analysis – this is a strategy widely used in medicine and implemented by some academic journals in Political Science and Psychology [14] (MIGUEL et al., 2014). Preregistration is a public document where the researcher specifies variables, data processing procedures, and analysis techniques (OLKEN, 2015). Preregistration aims to diminish publication bias, prevent the display of selective research results and reinforce its credibility (CASEY et al., 2012). Lastly, public data and materials enable other researchers to reproduce the research results, test remaining hypotheses and identify muddy results or even frauds (MIGUEL et al., 2014; MUFANO et al., 2017; CHRISTENSEN and MIGUEL, 2018).

The TIER Protocol fits into the third practice. More specifically, in the documentation that precedes data and replication materials' disclosure. The adoption of replicability policies by high impact factor journals pushes researchers to systematically document their research files. Despite demand, the materials made available are substandard. Stockemer, Koehler, and Lenz (2018) analyzed all the papers published in 2015 in the three main political behavior journals[15] and found that for 25% of them the code was so poorly organized that replication was impossible.

Well-organized replication materials enhance the expansion of related studies and enable more accessible

14 According to research by the authors, journals that have adopted policies towards this type of procedure are: *Journal of Experimental Political Science, Management Science,* and *Psychological Science.* See: Miguel et al. (2014), p. 30.

15 The authors analyzed all the papers present in the 2015 volumes of the following academic journals: *Electoral Studies*, *Party Politics*, and *Journal of Elections, Public Opinion & Parties*.

analyses (ALVAREZ et al., 2018). At the most basic level of preparation towards becoming a transparent researcher, it is recommended to adopt some sort of protocol for data organization to have reusable and humanly understandable data (FIGUEIREDO et al., 2019; OSINSKI, 2019).

Alvarez, Key, and Nuñez (2018) give some recommendations to increase data organization, clarity, and usage. The authors provide tips on how to archive materials related to the research in a useful and secure way, such as how to produce ReadMe files that are helpful for future replications, clues on how the files in replication materials should be named, as well as indicating how to report software (used or created by the author), scripts, and outputs generated by the research. Table 1 summarizes the authors' suggestions.

Table 1 - Recommendations for creating replication materials (Alvarez et al., 2018)

Processes	Recommendations
Archive	Replication materials must be stored in permanent files that will be available for long periods. Currently, the most trustworthy are those with institutional guarantees, whether in consortiums or universities; Authors must save their files so that they can be used in the future (e.g. comma-separated values file), rather than software-specific formats.
ReadMe files	The file must be sufficiently clear and detailed, but not overly. Some items must be included, such as (a) reference to the paper or associated publication; (b) a brief explanation of the files and the types of files included, e.g., original data, processed data, scripts, etc.; (c) an indication of the order the scripts should be executed; (d) a software list, software packages, and the operating system used to produce the results in the paper; (e) if uncommon extensions were used, the author must indicate how to proceed with those files.
File names	The files' names must be easy to comprehend and give information on their content. This is especially important for replication materials that include several scripts and data files. In case some files must be used in a certain order, they must be named in a way that includes the order (e.g., "1_ProcessedData, "2_Estimation", etc.).
User-created Software and Packages	When the author has created a statistical package, these must be (ideally) archived in a stable location, such as CRAN for R. If possible, the author should include a copy of the software or package in the replication materials.
Operating-system compatibility	Ideally, authors must ensure that their replication materials work in the most common operating systems. If it is not possible, they should indicate in which system they do work.
Scripts	(a) should contain a brief description of what they do, what they depend on, what packages are required, and what are the outputs; (b) lines without specification of what they do must be avoided, given that this indication helps in identifying specific areas of data management and processing and identifying possible errors or problems; (c) there should be good planning and recoding the data in the middle of the analysis should be avoided, unless extremely necessary.
Outputs	Replication materials should contain tables, graphs, and figures in the sequence that they appear in the paper.
Excluded outputs	Usually, replication materials have scripts or data that were not used in the paper or were linked to an online appendix. In this case, the author should differentiate them from the ones in the paper.
Intermediate outputs	Sometimes papers present intermediate outputs that lead to something essential in the research. In that case, they should be present in the replication materials.
Random sequences	For simulation or estimation strategies studies that use randomization, authors must always include the random sequence of numbers used to produce the results in the paper.
Directories and Folders	The directories' paths must be easily identifiable in the script, facilitating alteration by the users. It recommended that the directory is added only at the beginning, as modifying them throughout the script can create confusion.
Computing time	Authors must indicate (in the code and ReadMe file) if any script takes too long to calculate. This is important so that the user is not faced with a long estimation and believes they are having a problem.
Parallel Computing	More and more, authors are using parallel computing in their scripts to speed up the computational process. Unfortunately, this does not work on every computer. Therefore, they should indicate at which point in the script the strategy was used.
. Warnings and errors	It is not uncommon to find scripts that produce errors or warnings – that are usually unrelated to the inadequate use of packages or incompatibility with the operating system. In this case, the authors should indicate the reasons why these are no cause for concern.

Source: Alvarez et al. (2018), p. 425-426.

The suggestions by Alvarez et al. (2018) are, undoubtedly, a path towards producing more transparent and replicable research. Despite this, the number of steps required and the lack of more specific guidelines on implementing them make the work more difficult. Consequently, we defend adopting the TIER Protocol as a tool for the organization of replication materials.

We believe that the adoption and wide circulation of the TIER Protocol offers a clear benefit to the academic community. By providing a way to standardize replication materials, researchers in different fields intuitively know what they should find in each folder. In other words, the organization of replication materials will become more organic for researchers and more intuitive for replicators. Consequently, the TIER Protocol emerges as a less tortuous path to adopting the replicability standard.

These benefits are deeply rooted in the three dimensions of replicability: substantive, pedagogical, and transparent. In the substantive dimension, the replication standard allows research to be falsifiable and, thus, for science to move faster (PARANHOS et al., 2012; PARANHOS et al., 2013). Similarly, the TIER Protocol's adoption enabled the papers by Ball's students to be evaluated more cautiously and for students to received more precise comments on the problems faced in the production of the paper. For researchers who have decided to adopt the protocol, the organization strategy allows their papers to receive a better review by their peers, whether colleagues or journal reviewers.

Regarding the pedagogical dimension, the replication standard allows students to be initiated in the world of research and data analysis by having contact with real-world data and problems (KING, 1995). More recently, Ball's students solve *soup-to-nuts* exercises (more on that below), leading them to implement the protocol and go through the whole empirical research process with statistical data, from data collection to writing the report with statistical analyses. In like manner, the TIER hub in Brazil, based in Recife, has begun the production of workshops for undergraduate students in Political Science and International Relations, intending to disseminate the use of the protocol and get them initiated in empirical research.

Lastly, in the transparency dimension, the replication standard allows procedures and data to be publicly available to all researchers that are interested in the theme (KING, 1995; PARANHOS et al., 2013; FIGUEIREDO, 2019). The protocol enables that the availability of necessary information for replication is humanly intelligible and accessible so that any researcher can replicate the study's results without assistance from the author. For Ball's students, it enabled that the results were replicated without their presence. For science, it is the way that conflicting studies, discussions, and hypotheses advance towards producing knowledge.

Therefore, we believe that adopting the protocol can be a useful tool in the path of social scientists who aim to reach higher levels of transparency in their research, aiding not only those who are finding out the value of transparency now, but offering a clear map on how to proceed for those who were interested, but did not know how. In the next section, we present a *soup-to-nuts* exercise for replication, which introduces the protocol's application.

5. Exercise for the Application of the Protocol

In this section, we present the first *soup-to-nuts* exercise developed by the TIER initiative, to demonstrate the protocol's applicability (BALL, 2018).

The *soup-to-nuts* exercises were designed initially to be used in Introduction to Statistics of Introduction to Data Analysis classes[16]. In these activities, the aim is to lead the students to conduct a full empirical research process, from downloading a dataset through treatment, data analysis, and delivery of results. The steps' description is detailed and originally set to be reviewed in the classroom and then concluded by the students at home. In this activity, the students must build their folder hierarchy at the outset of the project working, therefore, with the TIER Protocol throughout the whole journey, given that they will be required to hand in the results with the complete and organized documentation. In addition to facilitating the professor's job when grading the exercises, as we have mentioned in Richard Ball's experience, the protocol helps students in the workflows, thus being a powerful example of this organizational methodology's potential. Therefore, the soup-to-nuts exercises are proper tutorials for students to work with data analysis in a transparent and replicable manner.

This paper will use a shortened version of the "Alcohol Consumption in American Universities" exercise as an example of the protocol[17]. We recommend that you dedicate some time to elaborate it, understand how the protocol contributes to the workflow, and facilitate the replication of a study. The complete version in Portuguese can be found in the OSF site[18] and the original version on the TIER website[19]. The solved exercise, with the report model and scripts are also available in the addresses indicated. The version presented below should be solved in R.

16 See: https://www.projecttier.org/tier-classroom/soup-nuts-exercises/#shorter-exercises-for-teaching-transparency-and-reproducibility . Last accessed on 20 March 2020.

17 Ball (2018) is the author of this exercise's original version, and Jenna Krall is credited as collaborator. The translation to Portuguese was done by members of the TIER hub Recife. This shortened version is by the authors of this paper, with authorization by the author of the original.

18 Complete version of the exercise in Portuguese: https://osf.io/mc2e9/?view_only=32e9fa640c83422ba7d6e9ab91e65743. Last accessed on 06 October 2020.

19 Complete original: https://www.projecttier.org/tier-classroom/soup-nuts-exercises/#shorter-exercises-for-teaching-transparency-and-reproducibility . Last accessed on 10 October 2020.

Replication Activity (TIER Protocol): Alcohol Consumption in American Universities

In this activity, you will explore if alcohol consumption is lower among students who live in housing where the practice is forbidden in relation to those who live in accommodation where the rule does not exist. The data comes from a study conducted in 2001 in American universities, by the *Harvard School of Public Health*. This activity will require you to use the data from that research to create some bar graphs and to answer a series of questions on their interpretation, comparing those who can drink in their housing and those who should not. Results to be handed in will not only be the report with graphs and answers but all the material used for processing and analysis conducted in this activity, following the documentation indicated.

I. PREPARATION

IA. Create a folder hierarchy to keep and organize your work.

Create a folder entitled **AlcoholExercise/.**

Inside **AlcoholExercise/**, create the following folders and subfolders:

- **Data/** (a subfolder of **AlcoholExercise/**)
  - **OriginalData/** (a subfolder of **Data/**)
  - **AnalysisData/** (uma subpasta **Data/**)
- ***Scripts/*** (a subfolder of **AlcoholExercise/**)
- ***Graphs/*** (a subfolder of **AlcoholExercise/**)

Save this folder hierarchy in the location you selected to save your work for this exercise.

I.B Install the packages

Two packages will be useful in this activity. Install them before beginning your work.

- *haven* (to read a dataset in *.dta* format)
- *tidyverse* (to facilitate processing procedures and creation of graphs)

II. DOWNLOAD AND EXAMINE THE DATABASE AND THE CODEBOOK YOU WILL USE IN THIS EXERCISE.

II.A. Find and explore the website where the data for this exercise are available.

The research data are archived at the *Inter-University Consortium for Political and Social Research* (ICPSR).

Go to the ICPSR's website (www.icpsr.umich.edu) and look for *Harvard School of Public Health College Alcohol Study, 2001*.

When you have found the page for this study, read the information provided.

II.B. Download the dataset and codebook.

- Before downloading the dataset, you will need to create an account at the ICPSR's website.

Once that is done, log in and go to the study's page: https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/4291

Click on the download button. You will see a menu with formats to choose from, including SAS, SPSS, Stata, and ASCII. In this menu, select Stata (with the *haven* package, R will read the *.dta* format).

The dataset's file will be downloaded in a zipped folder, *ICPSR_04291-V2.zip*.

When you extract the files, you will see that there are several documents, including a subfolder named **DS0001**. For this exercise, you will only use two files that came in the zipped folder and both are in the **DS0001** subfolder:

- *04291-0001-Data.dta*. This file contains the research data in *.dta* format. This file will be referred to as "the original dataset" in this exercise.
- *04291-0001-Codebook.pdf*. This is the codebook for the study used. It will be referred to as "the metadata for the original dataset".

You must save copies of these two files in the folder **OriginalData/**.

II. C. Examine the codebook and the dataset.

Examine the original codebook and dataset to familiarize yourself with the information contained in them. For this exercise, you will need five variables for this study. In the original dataset, these variables are named *A6, B8, B9, C13,* and *G11*. For each of these variables, analyzing the codebook, answer the following questions (the answers to these questions do not need to be in the final report, but they are key for you to understand the nature of the data you will be dealing with):

- How is the variable defined? What does it represent?
- What class is the variable (continuous, ordinal, categorical, discrete)?
- Is it qualitative or quantitative?
    - If it is quantitative, find the mean, the standard-deviation, the three quartiles, and the minimum and maximum values.
    - If it is qualitative categorical, how many categories are there, what do they represent, and what is the proportion of observations per category? For each

categorical variable, decide if the categories are or not ranked.

III. WRITING THE SCRIPTS

Working directory conventions

For all *scripts*, adopt the following working directory conventions:

- When all your scripts are executed, R's working directory must be set to the **Scripts/** folder.
- Do not write commands in your scripts that change the working directory. For instance, avoid the *setwd()* function.
- When you need to open or save a file in folder other than **Scripts/**, specify the folder's location in question using a relative directory path.

For this exercise, you will write three scripts:

III.A. *treatment.R*

- The purpose of this script is to modify the original dataset (*04291-001-Data.dta*), to prepare it for the analysis you will conduct.
    - Modifications include creating a new variable set and excluding a few observations in the dataset.
- Save the new dataset in a file named *analysis.RData*.
    - This file will be called analysis dataset.

Creating the script

- Open RStudio.
- Open a new R script.
- At the top of the new script, write a comment indicating that R's working directory should be the **Scripts/** folder. For example:

    # When this script is executed, the working directory should be the **Scripts/** folder.
- Save this new script in your **Scripts/** folder with the name *treatment.R*.
    - To avoid losing work, constantly save the script as you write it.
- Load the *haven* and *tidyverse* packages.
- Open (import) the *04291-0001-Data.dta* dataset.
    - Use the *read_dta()* function in the *haven* package.
    - Since your working directory is the **Scripts/** folder, you need to use a relative directory path to specifiy that the *04291-0001-Data.dta* file must be imported from **Original-Data/**.
- For this activity, we want to consider only students who live on campus or in fraternities/sororities. Therefore, exclude all cases in which the student's residence is "*Off campus house/apt*" or "*other*". Also exclude all the cases where the variable representing the student's residence is NA.
- Create a variable called *drunk*, which is:
    - Equal to 0 if the student drank enough to get drunk fewer three times in the past 30 days.
    - Equal to 1 if the student drank enough to get drunk three times or more in the past 30 days.
    - Equal to NA if the variable indicating how often the student drank enough

to get drunk in the past 30 days is NA.

- Create a variable called *hsdrunk*, which is:
    - Equal to 0 if the student drank five or more drinks in a row on two or less occasions during their last year of High School
    - Equal to 1 if the student drank five or more drinks in a row on three or more occasions during their last year of High School
    - Equal to NA if the variable indicating how many times the student had five or more drinks in a row during their last year of High School is NA
- Create a variable called *free*, which is:
    - Equal to 0 if the student does not live in alcohol-free housing
    - Equal to 1 if the student lives in alcohol-free housing
    - Equal to NA if the variable indicating whether or not the student lives in alcohol-free housing is NA.
- Create a variable called *volfree*, which is:
    - Equal to 1 if free=1 (the student lives in alcohol-free housing) and the student requested to live in alcohol-free housing
    - Equal to 0 if free=1 (the student lives in alcohol-free housing) and the student was assigned to live in alcohol-free housing
    - Equal to NA in all other cases
- Create a variable called *housing*, which is:
    - Equal to 1 if free=0 (the student does not live in alcohol-free housing)
    - Equal to 2 if free=1 (the student lives in alcohol-free housing) and the student was assigned to live in alcohol-free housing
    - Equal to 3 if free=1 (the student lives in alcohol-free housing) and all on-campus housing is alcohol-free
    - Equal to 4 if free=1 (the student lives in alcohol-free housing) and the student requested to live in alcohol-free housing
    - Equal to NA in all other cases
- Exclude all variables other than the six you just created.
- Exclude all cases where the values for *hsdrunk, drunk,* or *housing* are NA.
- Save the new dataset in the **AnalysisData/** folder, naming the file *Analysis.RData*.
    - You save the dataset with the *save()* function.
    - Here you will also need to indicate a relative directory path to specify the AnalysisData/folder's location.

Remember to save treatment.R in your **Scripts/** folder.

III.B. *DataAppendix.R*

The DataAppendix.R script will contain commands that generate the information that will appear in the Data Appendix (more info below).

Creating the script

- Open RStudio.
- Open a new R script.
- Open the data analysis file, analysis.RData (which, after having written and run with with treatment.R, should be saved in the **AnalysisData/** folder).
- Generate the following outputs for each of the five variables in the analysis dataset:
    - The number of NAs and the total

number of observations (NAs included)
    - A frequency distribution and a percent frequency distribution
    - A bar graph that presents the percent frequency distribution.

Remember to write comments explaining which lines produce which information and graph. Each line that produces information or graph that is part of the Data Appendix must be preceded by a comment that describes what is expected from this command. For example, above a line that generates a table showing the frequency distribution of the *housing* variable, you should write a comment such as "The following line generates a table showing the frequency distribution of the *housing* variable".

Remember to save DataAppendix.R in your **Scripts/** folder.

III.C. *analysis.R*

The analysis.R script generates the bar graphs for your main analysis in this activity, using the data from your analysis dataset (analysis.RData).

Creating the script

- Open RStudio.
- Open a new R script.
- Open analysis.RData (which, after having written and run with with treatment.R, should be saved in the **AnalysisData/** folder).
- Generate a bar graph, with two bars, where:
    - One bar represents students living in alcohol-free housing
    - One bar represents students not alcohol-free housing
    - The height of each bar is equal to the proportion of students (of the student group that the bar represents) that drank enough to get drunk three or more times in the past 30 days.
    - Save this graph to the **Graphs/** folder (subfolder of **AlcoholExercise/**), with the name Figure1.jpg. (Use a relative directory path to specify the location of the **Graphs/** subfolder).
- Generate a bar graph, with two bars, where:
    - One bar represents students who live in alcohol-free housing because they requested it
    - One bar represents students who live alcohol free housing because they were assigned to it
    - the height of each bar is equal to the proportion of students (of the student group that the bar represents) who drank enough to get drunk three or more times in the last thirty days
    - Save this graph in the **Graphs/** folder (subfolder of **AlcoholExercise/**), with the name Figure2.jpg.
- Generate a bar graph, with four bars, where:
    - Each bar represents students in one of the four categories defined by the *housing* variable
    - The height of each bar is equal to the proportion of students (in the group represented by the bars) who drank enough to get drunk three times or more in the past 30 days
    - Save this graph in the **Graphs/** folder (subfolder of **AlcoholExercise/**), with the name Figure3.jpg.
- Generate two bar graphs (side-by-side), each with two bars, where:
    - One of the graphs represents only students who had five or more drinks in a

row on three or more occasions during their last year of High School
    - One of the graphs represents only students who had five or more drinks in a row on less than three occasions during their last year of High School
    - And, in each of these graphs,
        - One bar representing students who live in alcohol-free housing
        - One bar representing students who do not live in alcohol-free housing
        - The height of each bar is equal to the proportion of students (in the group represented by the bars) who drank enough to get drunk three or more times in the past 30 days
        - Save this graph in the **Graphs/** folder (subfolder of **AlcoholExercise/**), with the name Figure4.jpg.
- Generate two bar graphs (side-by-side), each with two bars, where:
    - One of the graphs represents only students who had five or more drinks in a row on three or more occasions during their last year of High School
    - One of the graphs represents only students who had five or more drinks in a row on less than three occasions during their last year of High School
    - And, in each of these graphs,
        - One bar that represents students who live in alcohol-free housing because they requested it
        - One bar that represents students who live in alcohol-free housing because they were assigned to it
        - The height of each bar is equal to the proportion of students (in the group represented by the bars) who drank enough to get drunk three or more times in the past 30 days
        - Save this graph in the **Graphs/** folder (subfolder of **AlcoholExercise/**), with the name Figure5.jpg.
- Generate two bar graphs (side-by-side), each with four bars, where:
    - One of the graphs represents only students who had five or more drinks in a row on three or more occasions during their last year of High School
    - One of the graphs represents only students who had five or more drinks in a row on less than three occasions during their last year of High School
    - And, in each of these graphs,
        - Each bar represents students in one of the four categories defined by the *housing* variable
        - The height of each bar is equal to the proportion of students (in the group represented by the bars) who drank enough to get drunk three or more times in the past 30 days
        - Save this graph in the **Graphs/** folder (a subfolder of **AlcoholExercise/**), with the name Figure6.jpg.

Write comments indicating which lines produced

which graphs. For instance, before the line that generates Figure 4, write a comment such as "The following command line generates Figure 4".

Remember to save analysis.R in your **Scripts&Relatorio/** folder.

IV. THE REPORT

The report you must hand in for this activity will consist of presenting the figures generated and the answers to a series of questions that ask you to interpret the bar graphs you created. Save to *.pdf*.

Present the Figures and write a caption under each graph that indicates the number and an informative title. For instance, an appropriate caption for Figure 1 would be: "Figure 1: Rates of Heavy Drinking in Alcohol-Free vs. Not Alcohol-Free Housing".

Questions

1) Describe the sample in the AnalysisDataset.dta file that you created. What is the unit of analysis? How many observations are there? Describe the method used to create the sample and what criteria were considered to include or exclude groups of individuals.

Note that this question is about your cleaned and processed dataset, rather than the original dataset that you downloaded from ICPSR. However, to answer, you will need to turn to the original dataset's codebook.

2) Explain in your own words what Figure 1 shows. Was it what you expected? If not, how is Figure 1 different from what you expected?

3) Explain in your own words what Figure 2 shows. Does what you see in Figure 2 somehow clarifies what you saw in Figure 2? Explain.

4) How many individuals are represented in Figure 1? (That is, how many observations were used to generate the figure?). And how many individuals are represented in Figure 2? If the number of individuals represented in Figures 1 and 2 is not the same, explain: are some individuals shown in Figure 1 not shown in Figure 2? If so, which individuals were these? And are some individuals shown in Figure 2 not present in Figure 1? If so, which ones?

5) Considering Figure 3, which sections contain information already presented in Figures 1 and 2? Which sections show new information? Summarize in your words what Figure 3 presents.

6) Compare Figure 4 to Figure 1. Explain which new information is given by Figure 4 and give an interpretation for that information.

7) Compare Figure 5 to Figure 2. Explain which new information is given by Figure 5 and interpret it.

8) Compare Figure 6 to Figure 3. Explain which new information is given by Figure 6 and interpret it.

9) Of the figures you created, what general conclusions can be drawn regarding the factors associated with alcohol consumption among college students?

V. THE DATA APPENDIX

The Data Appendix should consist of multiple sections, one for each variable of the analysis dataset. Save it to *.pdf* as well.

For each variable, the section must have:

- The name of the variable.
- The number of NAs in the NA/Total format (number of NAs/total number of observations).
- A definition of the variable.

- Possible values for the variable.
- An explanation of what each value for the variable represents (codification of values).
- A table presenting the distributions of the frequency and percent frequency. If the categorical variable is ranked, that table must also present the cumulative percent frequency distribution.
- A bar graph that illustrates the distribution of the percentage frequency.

VI. THE READ ME FILE

The Read Me file is a document that gives information on the electronic files that you organized to document the work in this analysis. It should consist of two main sections:

> The first section should present a map of all the files included in the replication documentation and the folders and folder where they are.
>
> The second section should offer step-by-step instructions, explaining how to the documentation to (i) replicate all the necessary data processing to transform the original dataset into the analysis dataset, (ii) generate all the output you present in the Data Appendix, and (iii) reproduce the six bar graphs that you created.

These instructions must be written in clear language and detailed enough for someone unfamiliar with this exercise to follow and reproduce everything you have done. Save to *.pdf*.

VII. ELECTRONIC DOCUMENTATION

Once the exercise is finished, the files used and created must be allocated in the folder hierarchy in the following manner:

AlcoholExercise/
ReadMe.pdf
Report.pdf
Data/ (subfolder of **AlcoholExercise/**)
OriginalData (subfolder of **Data/**)
04291-0001-Data.dta
04291-0001-Codebook.pdf
AnalysisData (subfolder of **Data/**)
AnalysisDataset.dta
Scripts/ (subfolder of **AlcoholExercise/**)
treatment.R
DataAppendix.R
analysis.R
Graphs/ (subfolder of **AlcoholExercise/**)
Figure1.jpg
Figure2. jpg
Figure3. jpg
Figure4. jpg
Figure5. jpg
Figure6. jpg

Do not leave anything else to these folders

- No folders in addition to these.
- No files in addition to these.

Test your files to see if they run

- Find a computer other than the one where you conducted this activity and copy the whole folder into it, with all its contents.
- Go through the following steps, without changing the order of the folders or files:
  - Start R;
  - Define **Scripts/** as your working directory
  - Run treatment.R and, after that, check if the file AnalysisDataset.dta that you have in the **AnalysisData/** folder was altered by a new copy recently created (check the date and

time of modification of the file in the folder)

- Run DataAppendix.R and check if it generates the expected outputs. Also check of the graphs in the **Graphs/** folder were updated by new, recent versions.

Run analysis.R and check if the six graph files located in the **Graphs/** folder were updated.

6. Other TIER activities

Intending to promote the importance of transparency and disseminate the values of ethics in science, the TIER Project develops some promotion activities and events for its protocol and good scientific practices. All details can be found at their web address[20]. The TIER Project promotes workshops for faculty members and researchers to debate practices towards transparency and replicability in science and present to them the most recent ideas on the protocol (and gather comments and impressions)[21]. These workshops are excellent interuniversity and international cooperation opportunities for good scientific practices, bringing together researchers from different Social Sciences and different countries.

There is also the Fellowship program for young professors and researchers engaged in quantitative methods teaching and training with a transparent and replicable view. The scholarships last one academic year and seek to encourage the recipient to develop and disseminate transparent research teaching methods – such as *soup-to-nuts* exercises and workshops. Also, in February and March 2020, TIER promoted a series of conferences in webcast format. Each week, a guest would speak on their most recent developments and ideas on transparency and replicability. Among these "transparency leaders" were Gary King, Dorothy Bishop, and Scott Long. The presentations are available on TIER's website[22].

7. Conclusions

The principles and practices of transparency have become the gold-standard of empirical science. Initiatives to disseminate these ideas and present tools to increase transparency in research and scientific production have become more and more relevant. Replication materials gained strength as elements that compose finished research, and its publication and sharing have become a requirement for publication in several specialized journals, whether in the United States or, incipiently, in Brazil.

In this paper, we aimed to discuss ways to increase transparency in empirical research in the Social Sciences, mainly through documentation of replication materials. We presented, throughout the paper, the development of the TIER Protocol and its framing in the three dimensions of the replication standard (substantive, pedagogical, and transparency), since the protocol enables young researchers to be initiated in the world of empirical research through statistical data and data analysis, as well as increases the level of transparency of the works that adopt it. In addition, we presented the organizational structure of the protocol, the files included, and the folder hierarchy that follows the documentation suggested by the creators. We also presented a shortened version of a *soup-to-nuts* exercise to implement the protocol, ideal for application in Introduction to Statistics and Data Analysis classes. We hope

20 See: https://www.projecttier.org/ . Last accessed on 20 March 2020.

21 See: https://www.projecttier.org/fellowships-and-workshops/fall-2019-faculty-development-workshop/ . Last accessed on 20 March 2020.

22 See: https://www.projecttier.org/fellowships-and-workshops/weekly-webcast-leaders-research-transparency/ . Last accessed on 20 March 2020.

that the students who are introduced to these analyses already within a transparency and replicability paradigm have an easier time learning and developing them and carrying with them these values of good science.

Thus, we set about indicating a less tortuous path so that researchers can reach higher levels of transparency in their production. Also, we hope to have disseminated and presented the protocol in a simple and didactic manner. In this way, we hope to contribute with the transparency and replicability agenda, which has been growing in the Social Sciences as a whole, and in Brazilian Political Science in particular.

BIBLIOGRAPHICAL REFERENCES

ALVAREZ, R. M; KEY, E. M; NÚÑEZ, L. (2018). "Research replication: Practical considerations". *PS: Political Science & Politics,* 51(2), 422-426.

BALL, R.; MEDEIROS, N. (2012). "Teaching Integrity in Empirical Research: A Protocol for Documenting Data Management and Analysis". *The Journal of Economic Education*, 43(2), 182-189.

BALL, R. (2018). Animal House in Alcohol-Free Dorms? TIER Project. Disponível em: https://www.projecttier.org/tier-classroom/soup-nuts-exercises/#shorter-exercises-for-teaching-transparency-and-reproducibility. Acesso em 06 de outubro de 2020.

CASEY, K; GLENNERSTER, R; MIGUEL, E. (2012). "Reshaping institutions: Evidence on aid impacts using a preanalysis plan". *The Quarterly Journal of Economics*, 127(4), 1755-1812.

CHRISTENSEN, G.; SODERBERG, C. (2016). "Manual of best practices in transparent social science research". *Retrieved April*, 2, 2017.

CHRISTENSEN, G.; MIGUEL, E. (2018). "Transparency, reproducibility, and the credibility of economics research". *Journal of Economic Literature*, 56(3), 920-80.

DE LA GUARDIA, F. H.; STURDY, J. (2019). "Best Practices for Transparent, Reproducible, and Ethical Research". *IDB Technical Note* ; 1635.

FIGUEIREDO FILHO, D.; LINS, R.; DOMINGOS, A.; JANZ, N.; SILVA, L. (2019). "Seven Reasons Why: A User's Guide to Transparency and Reproducibility". *Brazilian Political Science Review*, 13(2).

JANZ, N. (2016). "Bringing the gold standard into the classroom: replication in university teaching". *International Studies Perspectives*, 17(4), 392-407.

KEY, Ellen M. (2016). "How are we doing? Data access and replication in political science. PS: Political Science & Politics, 2016, 49.2: 268-272

KING, G. (1995). "Replication, replication". *PS: Political Science & Politics*, 28(3), 444-452.

MIGUEL, E., CAMERER, C., CASEY, K., COHEN, J., ESTERLING, K. M., GERBER, A., LAITIN, D. (2014). "Promoting transparency in social science research". *Science*, 343(6166), 30-31.

MUNAFÒ, M. R., NOSEK, B. A., BISHOP, D. V., BUTTON, K. S., CHAMBERS, C. D., DU SERT, N. P., IOANNIDIS, J. P. (2017). "A manifesto for reproducible science". *Nature human behaviour*, 1(1), 1-9.

OLKEN, B. A. (2015). "Promises and perils of pre-analysis plans". *Journal of Economic Perspectives,* 29(3), 61-80.

OSINSKI, L. (2019). "PROOF course Open Science: The new default in science, 01-10-2019". *Presentation at Eindhoven University of Technology*. Disponível em: <https://www.projecttier.org/tier-classroom/course-materials/#modal230>. Acesso em 23 mar. 2020.

PARANHOS, R., FIGUEIREDO FILHO, D. B., da Rocha, E. C., da Silva Jr, J. A., & SANTOS, M. L. W. D. (2012). "Levando Gary King a sério: desenhos de pesquisa em Ciência Política". *Revista Eletrônica de Ciência Política*, 3(1-2).

PARANHOS, R., FIGUEIREDO FILHO, D. B., da ROCHA, E. C., CARMO, E. F. (2013). "A importância da replicabilidade na ciência política: o caso do SIGOBR". *Revista Política Hoje*, 22(2), 213-229.

PRZEWORSKI, A., SALOMON, F. (1988). "The Art of Writing Proposals: Some candid suggestions for Applicants to Social Science Research Council". *SSRC: New York*.