Relatório Técnico

Documento:

Relatório Técnico - Replicação dos Resultados da PeNSE2019

Data de emissão:

13/11/2023

Elaborado por:

Juliana Magalhães Rosa
Departamento de Estatística (EST)
Universidade de Brasília (UnB)

Sumário

1 Introdução

A Fundação Escola Nacional de Administração Pública (Enap), é um órgão do Governo Federal que visa a melhoria do serviço público a partir da oferta de formações e aperfeiçoamentos em Administração Pública aos servidores públicos federais. Sendo assim, cabe à Enap o desenvolvimento de pesquisas científicas/ tecnológicas e de novos produtos ou serviços relacionados à tecnologia de gestão.

Dentro da Enap, a Coordenação Geral de Ciência de Dados está destinada à produção de conhecimento a partir de dados gerados e armazenados pelo Governo Federal. Esse setor é, portanto, responsável pelo processo de carregar, transformar e analisar dados para extrair informações e tirar conclusões acerca de assuntos relevantes para a Gestão Pública.

Em agosto de 2023 surgiu uma demanda relacionada à Pesquisa Nacional de Saúde do Escolar (PeNSE), estudo realizado periodicamente pelo Instituto de Geografia e Estatística (IBGE). Essa demanda consistia em reproduzir os resultados obtidos pela edição de 2019 da PeNSE e apresentá-los em um formato mais utilizável para análise de dados, construção de painéis etc. Esse trabalho foi alocado a uma pesquisadora bolsista da Universidade de Brasília (UnB) por meio do Centro de Apoio ao Desenvolvimento Tecnológico - CDT/UnB.

2 Contextualização

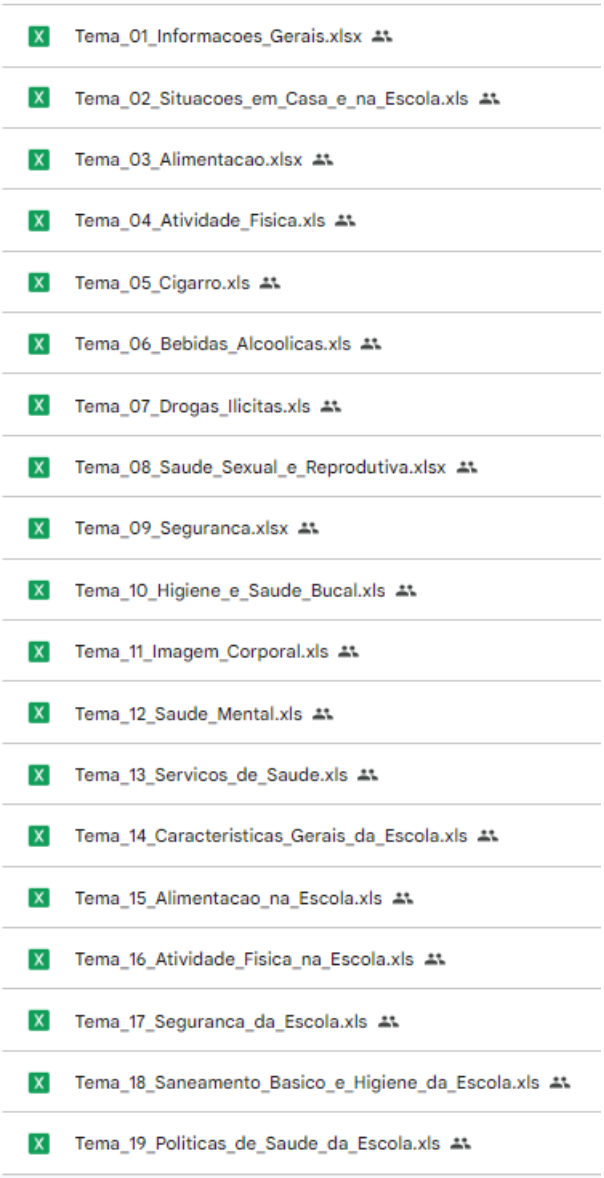
A PeNSE é uma pesquisa conduzida pelo IBGE desde 2009 com o intuito de identificar fatores de risco para a saúde de escolares no Brasil e subsidiar o desenvolvimento de políticas públicas voltadas para a promoção de saúde entre os estudantes da educação básica.

A edição de 2019 é a mais recente até então e teve como população alvo os escolares de 13 a 17 anos de escolas públicas e privadas. Além disso, assim como as edições anteriores, foi uma pesquisa por amostragem.

Com a finalização da pesquisa, o IBGE divulga os microdados coletados (com exceção de algumas variáveis que são omitidas por questão de sigilo), juntamente com os resultados obtidos, em forma de planilhas, relatórios e outros documentos. Os principais resultados consistem em estimativas de indicadores de percentuais e de totais dos escolares.

No caso da PeNSE 2019, essas estimativas foram apresentadas em 19 arquivos *Excel*, um para cada tema da pesquisa, conforme a Figura 1. Os cálculos desses valores foram replicados no presente trabalho, como parte da demanda da ENAP.

Figura 1: Planilhas com estimativas de indicadores calculadas na PeNSE 2019.

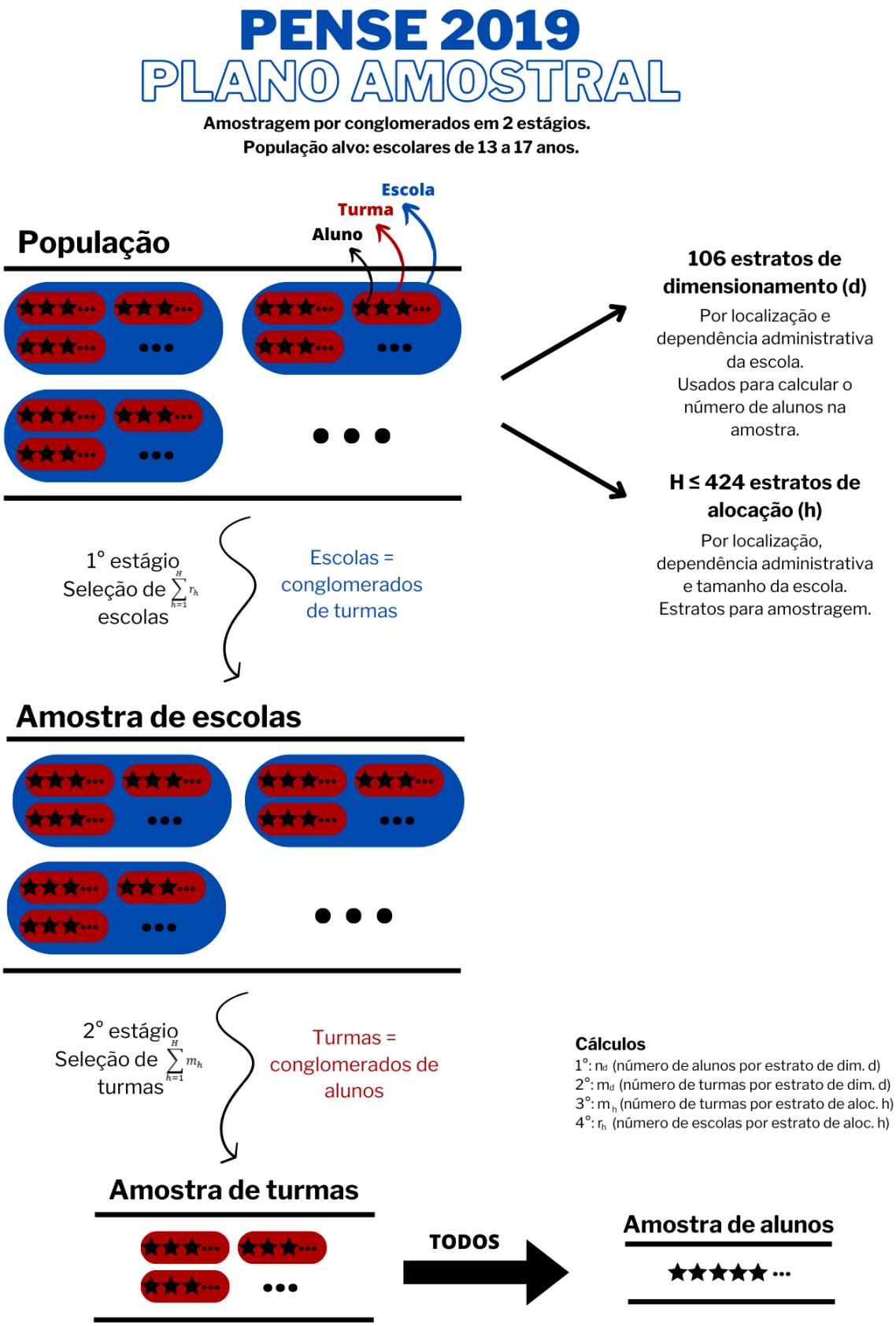


3 Desenvolvimento do Trabalho

3.1 Estudo do plano amostral

Para o entendimento do desenho amostral utilizado na pesquisa foi necessário o estudo de amostragem complexa e a leitura do relatório da PeNSE 2019. Além disso, nessa etapa do trabalho foi elaborada a Figura 2, a fim de apresentar esse plano amostral de maneira mais clara e visual.

Figura 2: Plano amostral da PeNSE 2019.



Levando em consideração esse tipo de amostragem e a fim de ilustrar alguns dos cálculos realizados, é apresentada abaixo a expressão para o estimador do total Y de uma variável y de interesse:

$$\hat{Y} = \sum_{h=1}^{H} \frac{1}{r_h} \sum_{i=1}^{r_h} \frac{\hat{Y_{hi}}}{p_{hi}} \tag{1}$$

onde:

H é o número total de estratos de alocação na população;

r_h é o número de escolas amostradas no estrato h;

$p_{hi} = \frac{M_{hi}}{M_h}$ é o tamanho relativo da i-ésima escola no estrato h na população;

M_{hi} é o número de turmas da i-ésima escola no estrato h na população;

M_h é o número total de turmas no estrato h na população;

$\hat{Y_{hi}} = \sum_{j=1}^{m_{hi}} w_{j/hi} y_{hij}$ é o estimador simples do total da variável y na i-ésima escola no estrato h;

m_{hi} é o número de turmas da i-ésima escola no estrato h na amostra;

$w_{j/hi} = \frac{M_{hi}}{m_{hi}}$ é o peso da j-ésima turma dado a seleção da i-ésima escola no estrato h;

e y_{hij} é o valor da variável y na j-ésima turma da i-ésima escola no estrato h.

3.2 Construção do código em R

Originalmente, o processamento dos dados e o cálculo dos indicadores da pesquisa foram realizados pelo IBGE no *software* SAS, mas os programas utilizados não são divulgados ao público. Assim, as replicações feitas neste trabalho foram calculadas no R, sendo este um *software* livre e de uso disseminado entre cientistas de dados, inclusive dentro da Enap.

Todo o código utilizado pode ser encontrado no repositório particular da pesquisadora e no repositório institucional da Enap. Para a reprodução deste trabalho, as pastas devem ser organizadas localmente da mesma forma em que se encontram dispostas no repositório

e os arquivos *Excel* com os dados devem ser inicialmente substituídos por versões vazias com os mesmos nomes.

O primeiro *script* a ser executado deve ser o **inicializacao.R**, seguido dos arquivos **funcoes.R** e **modelos.R**. Tendo feito isso, as pastas para cada tema podem ser acessada e seus *scripts* executados para cada arquivo *Excel* que se deseje gerar. Nos casos em que existe um *script* de ajuste dos dados, como nos temas 9 e 17, este deve ser executado antes do código que gera as tabelas.

3.2.1 Funções

Primeiramente, foram criadas cinco funções no R:

```
estima_total(dados, desenhos, nomes_vars)
```

a qual calcula as estimativas de total. Recebe os seguintes argumentos:

- $dados$: conjunto de dados;
- $desenhos$: desenho amostral criado a partir do pacote $survey$;
- $nome\_vars$: vetor com os nomes das variáveis desejadas na estimação.

```
estima_pct(dados, desenhos, nomes_vars)
```

a qual realiza a estimação de percentual e recebe os mesmos argumentos de $estima\_total()$.

```
tab_2vars(tt, indicador, var1, valor_var1, var2, valor_var2, filtro)
```

a qual organiza as estimativas em uma tabela de formato longo, para os casos em que **duas** variáveis foram utilizadas na estimação. Recebe os seguintes argumentos:

- tt: tabela de estimativas gerada por $estima\_total()$ ou $estima\_pct()$;
- $indicador$: nome do indicador (”Total de escolares”ou ”Percentual de escolares”);
- $var1$: nome da primeira variável utilizada na estimação;
- $valor\_var1$: vetor de valores que a primeira variável recebe (na ordem apropriada);
- $var2$: nome da segunda variável utilizada;
- $valor\_var2$: vetor de valores da segunda variável;

- $filtro$: vetor numérico de posições para filtrar a tabela e manter apenas as linhas desejadas (quando não se quer incluir todas as categorias de uma variável).

```
tab_3vars(tt, indicador, var1, valor_var1, var2, valor_var2, var3, valor_
    var3, filtro)
```

a qual gera uma tabela com as estimativas no formato longo, para os casos em que **três** variáveis foram utilizadas na estimação. Seus parâmetros seguem a mesma ideia daqueles da função $tab\_2vars()$.

```
tab_4vars(tt, indicador, var1, valor_var1, var2, valor_var2, var3, valor_
    var3, var4, valor_var4, filtro)
```

a qual forma uma tabela com as estimativas no formato longo, para os casos em que **quatro** variáveis foram utilizadas na estimação. Seus parâmetros seguem a mesma ideia daqueles das funções $tab\_2vars()$ e tab_3vars.

3.2.2 Modelos

Outras dez funções foram criadas e utilizam as funções já apresentadas anteriormente dentro de si. Foram chamadas de ”modelos”, por se basearem nos diferentes modelos de tabelas apresentadas nas planilhas do IBGE.

```
modelo1(DESENHO, VAR_COL, NOME_VAR_COL, VETOR_COL, FILTRO, fun_estima,
    fun_arruma)
```

que é o modelo em que o IBGE utiliza os grupos de idades e as regiões do Brasil nas linhas da tabela, com outras variáveis nas colunas. Recebe os seguintes argumentos:

- $DESENHO$: desenho amostral;
- VAR_COL: nome da variável que aparece nas colunas da tabela do IBGE, conforme está no conjunto de dados;
- $NOME\_VAR\_COL$: nome que será dado a essa variável na tabela de saída;
- $VETOR\_COL$: vetor com os valores dessa variável;
- $FILTRO$: vetor numérico com as posições das linhas que se quer manter na filtragem;

- fun_estima: nome da função usada para a estimação (”estima_total”ou ”estima_pct”);
- fun_arruma: nome da função usada para a organização da tabela em formato longo (”tab_2vars”, ”tab_3vars”ou ”tab_4vars”).

```
modelo1_2(DESENHO, VAR_COL, NOME_VAR_COL, VETOR_COL, FILTRO, VAR_LIN, NOM
    _VAR_LIN, VETOR_LIN, fun_estima, fun_arruma)
```

que é o modelo em que as linhas da tabela contêm informação sobre alguma variável a mais, como sexo do aluno ou dependência administrativa da escola. Possui os mesmos parâmetros do $modelo1()$ e mais alguns relacionados à nova variável:

- VAR_LIN: indicativo dessa coluna no conjunto de dados, na forma $dados\$variavel$;
- NOM_VAR_LIN: nome dessa variável na tabela de saída;
- $VETOR\_LIN$: vetor com os valores dessa variável.

```
modelo2(DESENHO, VAR_COL, NOME_VAR_COL, VETOR_COL, FILTRO, fun_estima,
    fun_arruma)
```

que é um modelo com as regiões brasileiras e as unidades da federação nas linhas da tabela. Recebe os mesmos parâmetros já apresentados.

```
modelo3(DESENHO, VAR_COL, NOME_VAR_COL, VETOR_COL, FILTRO, fun_estima,
    fun_arruma)
```

que é um modelo com as capitais das unidades da federação nas linhas da tabela. Recebe os mesmos parâmetros já apresentados.

```
modelo4(DESENHO, VAR_COL, NOME_VAR_COL, VETOR_COL, FILTRO, VAR_LIN, NOM_
    VAR_LIN, VETOR_LIN, fun_estima, fun_arruma)
```

que é um modelo com a raça dos alunos e alguma outra variável (como o sexo dos alunos ou a dependência administrativa das escolas) nas linhas da tabela. Recebe os mesmos parâmetros já apresentados.

```
modelo5(var, var_string, var_titulo, valor, filtragem)
```

é um modelo que chama os modelos 1, 2 e 3, já gerando os três tipos de tabelas como resultados. Mas serve para os casos em que as informações de sexo do aluno e dependên-

cia administrativa da escola estão nas colunas da tabela, além de uma outra variável a ser informada. Recebe como argumentos:

- var: indicativo dessa coluna no conjunto de dados, na forma $dados\$variavel$;
- var_string: o nome da variável como $string$;
- var_titulo: título que essa variável deve receber na tabela de saída;
- $valor$: vetor de valores dessa variável;
- $filtragem$: vetor numérico com posições para a filtragem.

```
modelo5_1(var, var_string, var_titulo, valor, filtragem, var_filtro,
    valor_filtro)
```

o qual é igual ao $modelo5()$, mas com uma filtragem a mais no desenho amostral. Os parâmetros a mais são:

- var_filtro: variável a ser filtrada;
- $valor\_filtro$: valor que se quer excluir.

```
modelo6(var, var_string, var_titulo, valor, filtragem)
```

o qual é igual ao $modelo5()$, mas inclui os valores -1, que representam abandono de questionário, nos cálculos. Esses valores geralmente são ignorados nas estimações, mas podem ser necessários a depender do contexto.

```
modelo7(var, var_string, var_titulo, valor, filtragem)
```

o qual é similar ao $modelo5()$, mas contendo apenas a dependência administrativa nas colunas da tabela (além da outra variável de interesse), não incluindo a variável sexo.

```
modelo8(var, var_string, var_titulo, valor, filtragem)
```

o qual é igual ao $modelo7()$, mas inclui os valores -1 nos cálculos.

3.2.3 Inicialização dos cálculos

Antes de aplicar as funções listadas para de fato calcular as estimativas dos 19 temas, foram necessários alguns procedimentos de inicialização.

1. Carregar pacotes no R.
    - $vroom$ para leitura dos dados;
    - $survey$ para criação do desenho amostral e cálculo das estimativas;
    - $dplyr$ para manipulação do conjunto de dados e organização das tabelas de saída;
    - $openxlsx$ para exportação das planilhas.
2. Ler os microdados da PeNSE 2019.
3. Desenhar o plano amostral no R.
4. Filtrar dos dados para a população-alvo (escolares de 13 a 17 anos).
5. Fazer ajustes nos dados (criação de novas variáveis, agregação de categorias de uma mesma variável, etc).

    Para mais detalhes, consultar o *script* **inicializacao.R**.

3.2.4 Geração e exportação das planilhas

As planilhas foram geradas a partir das funções e dos modelos explicados, tendo formato longo, conforme exemplo na Figura 3.

Figura 3: Exemplo de estimativas em formato longo.

	A	B	C	D	E	F	G	H	I	J	K
1	Indicador	Variavel1	Valor_Var1	Variavel2	Valor_Var2	Variavel3	Valor_Var3	Estimativa	Erro_Padrao	LI	LS
2	Total de esco	Idade	13 a 17 anos	Localização	Brasil	Sexo	Masculino	5843329,03	66632,3109	5712732,1	5973925,96
3	Total de esco	Idade	13 a 17 anos	Localização	Brasil	Sexo	Feminino	6006023,83	75576,8998	5857895,83	6154151,84
4	Total de esco	Idade	13 a 15 anos	Localização	Brasil	Sexo	Masculino	3777634,5	70180,3776	3640083,49	3915185,51
5	Total de esco	Idade	13 a 15 anos	Localização	Brasil	Sexo	Feminino	3882335,79	80858,7777	3723855,5	4040816,09
6	Total de esco	Idade	16 e 17 anos	Localização	Brasil	Sexo	Masculino	2059666,79	42587,4612	1976196,9	2143136,68
7	Total de esco	Idade	16 e 17 anos	Localização	Brasil	Sexo	Feminino	2117884,43	54411,2567	2011240,33	2224528,54
8	Total de esco	Idade	13 a 17 anos	Localização	Norte	Sexo	Masculino	600873,982	12354,4104	576659,783	625088,181

Além das estimativas pontuais, as saídas incluem os erros padrão calculados e os respectivos limites inferiores (LI) e limites superiores (LS) dos intervalos de confiança.

A exportação dessas planilhas é feita através do pacote $openxlsx$ do R e elas são separadas em 19 arquivos, um para cada tema da pesquisa, análogo aos resultados originais publicados pelo IBGE.

4 Considerações Finais

Ao longo do projeto, foi mantida comunicação com a equipe do IBGE responsável pela PeNSE 2019, a fim de buscar esclarecimentos e validações em relação ao plano amostral, aos dados e aos cálculos das estimativas. Em uma das reuniões realizadas nesse sentido, foi constatado que os microdados não são divulgados em sua totalidade. Por questão de confidencialidade, algumas informações são omitidas do público geral, conforme explicado na nota técnica 01/2022.

Sendo assim, é esperado que uma replicação dos resultados da pesquisa não atinja valores idênticos aos encontrados originalmente, sendo possível apenas chegar a aproximações. Além disso, a omissão de algumas colunas do conjunto de dados impossibilitou a reprodução de 60 planilhas [1], dentre as 642 montadas pelo IBGE.

Com exceção das 60 tabelas mencionadas, todos os demais indicadores divulgados pelo IBGE puderam ser reproduzidos de forma bem próxima. As variações em relação aos valores originais foram, em geral, de no máximo duas casas decimais. Apenas a planilha 15.16.2 apresentou uma diferença maior, mas por equívoco nos valores divulgados originalmente, conforme esclarecido em troca de e-mails com a equipe da PeNSE 2019.

De forma geral, foram bem-sucedidas a replicação dos resultados da PeNSE 2019, a documentação de todos os cálculos e processos realizados com esse fim, e a entrega desses resultados em formato longo, de forma a facilitar futura utilização dos indicadores para a construção de painéis, a realização de análises ou a divulgação dos valores através de outras ferramentas.

[1] Planilhas 14.2.1, 14.2.2, 14.2.3, 14.6.1, 14.6.2, 14.6.3, 14.7.1, 14.7.2, 14.7.3, 14.8.1, 14.8.2, 14.8.3, 15.2.1, 15.2.2, 15.2.3, 15.3.1, 15.3.2, 15.3.3, 15.4.1, 15.4.2, 15.4.3, 15.5.1, 15.5.2, 15.5.3, 15.6.1, 15.6.2, 15.6.3, 16.4.1, 16.4.2, 16.4.3, 16.5.1, 16.5.2, 16.5.3, 16.7.1, 16.7.2, 16.7.3, 16.9.1, 16.9.2, 16.9.3, 16.10.1, 16.10.2, 16.10.3, 16.11.1, 16.11.2, 16.11.3, 16.13.1, 16.13.2, 16.13.3, 16.14.1, 16.14.2, 16.14.3, 16.15.1, 16.15.2, 16.15.3, 18.2.1, 18.2.2, 18.2.3, 18.5.1, 18.5.2, 18.5.2.